DOCUMENT RESUME

ED 380 474 TM 022 775

AUTHOR	Longford, N. T.
TITLE	Model-Based Methods for Analysis of Data from 1990 NAEP Trial State Assessment. Research and Development Report.
INSTITUTION	Educational Testing Service, Princeton, N.J.
SPONS AGENCY	National Center for Education Statistics (ED), Washington, DC.
REPORT NO	ISBN-0-16-045452-2; NCES-95-696
PUB DATE	Jan 95
NOTE	87p.
AVAILABLE FROM	U.S. Government Printing Office, Superintendent of Documents, Mail Stop: SSOP, Washington, DC 20402-9328.
PUB TYPE	Reports - Evaluative/Feasibility (142)
EDRS PRICE	MF01/PC04 Plus Postage.
DESCRIPTORS	*Data Analysis; *Estimation (Mathematics); Models; National Surveys; Regression (Statistics); *Sampling; Statistical Bias; *Statistical Distributions
IDENTIFIERS	*Cluster Sampling; Jackknifing Technique; Mean (Statistics); National Assessment of Educational Progress; *Trial State Assessment (NAEP); Variance (Statistical); Weighting (Statistical)

ABSTRACT

Model-based methods for estimating the population mean in stratified clustered sampling are described. The importance of adjusting the weights is assessed by an approach considering the sampling variation of the adjusted weights and its (variance) components. The resulting estimators are more efficient than the jackknife estimators for a variety of datasets obtained from the 1990 Mathematics Trial State Assessment of the National Assessment of Educational Progress (NAEP). The methods can be extended to two-stage clustering. A general method for estimation of more complex population summaries, such as regression coefficients, is outlined. There are no distributional assumptions in model-based methods, apart from the normality of the sample means. Model-based methods use only the final adjusted weights; the replicate weights can be disposed of, thus radically reducing the size of the dataset and simplifying data handling procedures. The principal advantage of the model-based methods is in efficiency and small bias of the estimators of standard errors for the population mean. Contrary to theoretical claims, the NAEP operationally implemented jackknife estimator of the sampling variance is not unbiased. Eleven tables and 7 figures are included. (Contains 13 references.) (Author/SLD)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

NATIONAL CENTER FOR EDUCATION STATISTICS

Research and Development Report January 1995

Model-Based Methods for Analysis of Data From 1990 NAEP Trial State Assessment

U.S. DEPARTMENT OF EDUCATION
Office of Educational Research and Improvement
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

☑ This document has been reproduced as received from the person or organization originating it.

☐ Minor changes have been made to improve reproduction quality.

• Points of view or opinions stated in this document do not necessarily represent official OERI position or policy.

ED 380 474

TM022775

U.S. Department of Education
Office of Educational Research and Improvement NCES 95-696

NATIONAL CENTER FOR EDUCATION STATISTICS

Research and Development Report January 1995

Model-Based Methods for Analysis of Data From 1990 NAEP Trial State Assessment

N. T. Longford
Educational Testing Service, Princeton, NJ

U.S. Department of Education
Office of Educational Research and Improvement NCES 95-696

U.S. Department of Education
Richard W. Riley
Secretary

Office of Educational Research and Improvement
Sharon P. Robinson
Assistant Secretary

National Center for Education Statistics
Emerson J. Elliott
Commissioner

National Center for Education Statistics

"The purpose of the Center shall be to collect, analyze, and disseminate statistics and other data related to education in the United States and in other nations."—Section 406(b) of the General Education Provisions Act, as amended (20 U.S.C. 1221e–1).

January 1995

Contact:
Alex Sedlacek
(202) 219–1734

For sale by the U.S. Government Printing Office
Superintendent of Documents, Mail Stop: SSOP, Washington, DC 20402-9328
ISBN 0-16-045452-2

Foreword

The Research and Development (R&D) series of reports has been initiated

1) To share studies and research that are developmental in nature. The results of such studies may be revised as the work continues and additional data become available.

2) To share results that are, to some extent, on the "cutting edge" of methodological developments. Emerging analytical approaches and new computer software developments often permit new, and sometimes controversial analysis to be done. By participating in "frontier research," we hope to contribute to the resolution of issues and improved analysis.

3) To participate in discussions of emerging issues of interest to educational researchers, statisticians, and the Federal statistical community in general. Such reports may document workshops and symposiums sponsored by NCES that address methodological and analytical issues or may share and discuss issues regarding NCES practice, procedures, and standards.

The common theme in all these goals is that these reports present results or discussion that do not reach definitive conclusions at this point in time, either because the data are tentative, the methodology is new and developing, or the topic is one on which there are divergent views. Therefore the techniques and inferences made from the data are tentative and are subject to revision. To facilitate the process of closure on the issues, we invite comment, criticism, and alternatives to what we have done. Such responses should be addressed to:

Emerson Elliott
Commissioner
National Center for Education Statistics
555 New Jersey Ave. NW
Washington, D.C. 20208

Abstract

Model-based methods for analysis of surveys with stratified clustered design are discussed and applied to the 1990 NAEP Trial State Assessment. The principal advantage of the model-based methods is in statistical efficiency, and computational simplicity for regression analysis. Model-based methods dispense with the replicate weights which form a large part of the survey data.

Some key words: clustering, regression, stratification, variance components.

Contents

Foreword iii
Abstract iv
Acknowlegements viii

1 Introduction 1

2 The sampling design 2
2.1 Elements of the sampling design . . . 2
2.1.1 The sampling frame . . . 2
2.1.2 Selecting the schools . . . 2
2.1.3 Replicate groups – stratification . . . 3
2.1.4 Selecting students . . . 4
2.1.5 Sampling weights . . . 5
2.2 Proficiency scores . . . 5
2.3 Notation . . . 6

3 Model-based estimation of standard errors in stratified clustered sampling 8
3.1 Modelling features of the design . . . 8
3.1.1 Sampling weights . . . 8
3.1.2 Clustering . . . 9
3.1.3 Stratification . . . 10
3.1.4 Adjustment for non-response and poststratification . . . 10
3.1.5 Estimation of proficiency . . . 11
3.2 Randomness and conditioning in inference . . . 12
3.3 Jackknife . . . 13
3.4 Model-based methods . . . 14
3.4.1 The sampling variance of the mean . . . 15
3.4.2 Estimation of the variance components . . . 16
3.5 Examples . . . 19
3.5.1 Population mean . . . 20
3.5.2 Subpopulation means . . . 23
3.6 Adjustment of weights for nonresponse . . . 23
3.7 Association of weight adjustment and proficiency . . . 27
3.8 Multivariate outcomes . . . 28
3.9 Modelling approach . . . 28

4 Simulations 29
4.1 Estimators for subpopulations . . . 34

5 Smoothing techniques **36**

6 Regression with survey data **39**

6.1 Residual variance 42

6.2 Implementation 42

6.3 Regression with jackknife 43

6.4 Example 43

6.5 Multivariate and multilevel regression 43

7 Two-stage clustered sampling design **45**

7.1 Estimators of the variance components 50

7.2 Stratified two-stage clustered design 51

8 Summary **51**

9 Appendix. Data analysis with Splus **52**

9.1 Data input 52

9.2 A data summary 53

9.3 Jackknife 55

9.4 Model based estimation 59

9.5 Regression 67

References **76**

List of Tables

1 Clustering structure of the New Jersey sample. 4
2 Coefficients for the within-group sums of squares v_B and v_2. 19
3 Jackknife analysis. Estimation of the population mean of proficiency scores. 20
4 Model-based estimators of standard error of the weighted mean. 21
5 Model-based estimators of standard error of the weighted mean. Estimation of the population mean of proficiency scores for Oklahoma. 22
6 Jackknife and model-based estimates for a selected set of subpopulations; New Jersey. 24
7 Summary of simulation of model-based estimators. 31
8 Distribution of Hispanic students in the New Jersey sample. 35
9 Summary of simulation of model-based estimators for Hispanic students. 36
10 Regression analysis using the jackknife and model-based methods; New Jersey data. 44
11 Regression analysis of the proficiency on a constructed variable using the model-based method; New Jersey data. 45

List of Figures

1 Weighted systematic sampling of schools. 3
2 Adjusted weights for New Jersey. 5
3 Final weights and proficiency scores for the New Jersey sample. 7
4 Student-level weight adjustment factors for Oklahoma. 25
5 Association of cluster- and student-level weight adjustment factors with proficiency; New Jersey. 28
6 Comparison of the estimated and simulated within-cluster log-variances. 30
7 Pairwise plots of the simulated estimates of sampling variance using jackknife and methods CI, CII, CIII, and REML. 33

Acknowledgements

Final report on the research funded by the grant 999B20004 from the National Center for Educational Statistics. Discussions with and reviews of earlier drafts by Cliff Clogg, Paul Holland, Frank Jenkins, Eugene Johnson, John Mazzeo, Keith Rust, and Neal Thomas are acknowledged. Tom Jirele provided competent computing assistance.

1 Introduction

Just like other large scale surveys, those comprising the 1990 Math Trial State Assessment Program have a complex sampling design several features of which invalidate statistical analyses based on routinely adopted assumptions. A large part of this report is concerned with efficient estimation of the mean of proficiency for a population of students within a state, and of the standard error of such estimators, that take account of the salient features of the survey design.

We briefly summarize these features of the sampling design and indicate our approach. The students in the sample are associated with (unequal) sampling weights; they are clustered within schools, and schools are assigned to groups which for the purposes of analysis are regarded as strata. Each stratum is represented in the sample by a small number of schools (two for most strata), and most selected schools are represented by 20 – 30 students. The sampling procedures at the school level (selection of schools) and within a *selected* school are conditionally independent given the selected schools.

The *design weight* is the reciprocal of the probability, intended by the sampling design, of selecting a given student into the survey. It is the product of the (intended) probability of selecting the school, and of the (intended) conditional probability of selecting the student given selection of his/her school. Nonresponse of schools and individual students is compensated by adjusting the weights. This adjustment is not precise in the sense that the adjusted weights are not reciprocally proportional to the conditional probabilities of inclusion in the survey, given that the population of interest contains non-respondents.

The adjustment depends on the sample drawn, and it is therefore meaningful to regard the adjusted weights as *random* variables. For each subject we consider the (unknown) average adjustment over all the samples that could have been drawn, and the actual adjustment calculated based on the drawn sample and its pattern of non-response. The variation over the hypothetical samples of the normalized difference, or the log-ratio, of these two quantities is an informative summary of the weight adjustment.

The outcome variable is the proficiency score. This score is defined by reference to a model relating the student's ability and item characteristics to the probability of correct response, see Mislevy (1984) for details. The proficiency score is itself estimated from the students' responses to cognitive items. A set of five *exchangeable* estimates of the proficiency score, called the *plausible values*, are defined for each student. In addition to the general proficiency scale for mathematics subscales are defined for five content areas within the domain of mathematics. The report focuses on the general proficiency scores, but the methods presented are also applicable to the subscores. The methods used to obtain the proficiency scores and their actual values are accepted without criticism.

The computational algorithms described in this report are implemented in the statistical package Splus (Becker, Chambers, and Wilks, 1988), and some of them are documented in the Appendix. The principal advantages of Splus over statistical software established in quantitative educational research (such as SPSS, GLIM, or SAS) are flexibility (in both interactive and batch modes), ease of development of complex programs (functions), high quality graphics, and integrity of the environment generated by the defined data, functions, and other objects.

2 The sampling design

The original intention was to draw from the population of eighth-graders in each participating state or territory a sample of 105 schools and 30 students from each selected school that has more than 35 eighth-graders, and all the students from schools with fewer than 35 eighth-graders. In some states a small number of schools were included in the sample with certainty, and a larger number of students was drawn from each of these 'certainty' schools. For states (or territories) with fewer than 105 schools each school would be included in the sample. In states with an appreciable proportion of students in small schools ('small' meaning fewer than 20 eighth-graders), aggregate units (sets of schools) containing more than 20 students would be the units of sampling. Several factors intervened in this design, including non-cooperation of schools (school districts) and non-response of students, and incomplete and inaccurate information relevant to the sampling frame. Some states and territories (such as, Delaware, Guam, and Virgin Islands) have fewer than 105 schools. Allowing for non-response and small schools, it was expected that the sample for each state would comprise at least 2000 students from at least 100 schools.

2.1 Elements of the sampling design

2.1.1 The sampling frame

The sampling frame (the list of schools in the state) was constructed using several official sources, such as NCES Common Core of Data and Quality Education Data, Inc. The frame consisted of a list of all schools in the state that have eighth-grade students, the (estimated or exact) number of eighth-graders, or the exact number in a previous year, and the stratifying variables, defined for the school district or another administrative unit: urbanicity (city, suburban, and 'other'), median household income (grouped ordinal categories), and, where prevalent, minority enrollment (high enrollment of black and/or Hispanic students). See Koffler (1991) for details.

Schools with fewer than 20 eighth-grade students ('small' schools) were either attached to schools in their geographic proximity to form units with more than 20 students or aggregated into units with 20 or more students each.

2.1.2 Selecting the schools

A natural ordering of the strata was defined, combining urbanicity and minority enrollment. The schools were sorted in a 'serpentine' order, from the lowest median income to the highest in the first stratum, from the highest median income to the lowest in the next stratum, and so on.

In some of the states a small number of schools, C, were included in the sample with certainty. The rest of the schools are referred to as non-certainty schools. From the sorted list of non-certainty schools a systematic sample was drawn, with a random start, probability proportional to school enrollment, and step-length such as to ensure that K schools would be selected. For most states $C + K = 105$. This systematic sampling scheme is best illustrated as follows: The non-certainty schools are represented on a straight line by segments of lengths proportional to eighth-grade enrollment. A step-length s for a systematic sample

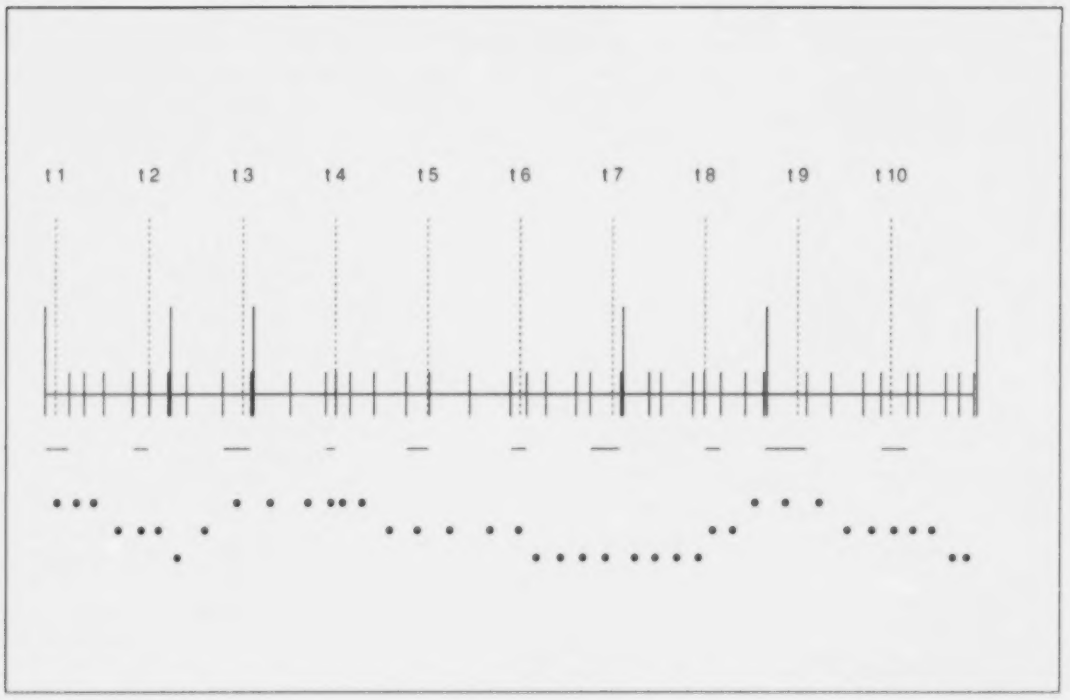

Figure 1: Weighted systematic sampling of schools.
Notes: The schools are delineated by short ticks, the groups by long ticks. The sampled points are indicated by long dotted ticks (points $t1, t2, \ldots$). Median income level (three categories) is indicated by asterisk. The segments of the selected schools (containing the sampled points) are underlined.

from this line is set, and the start is chosen as the point in distance t_1 from the origin, where t_1 is a draw from the uniform distribution on $(0, s)$. Further sampled points are $t_1 + ks$, $k = 1, 2, \ldots K - 1$, where K is the desired number of points; s is chosen so that Ks is the length of the segment corresponding to all the schools. The schools which correspond to the segments containing the selected points are included in the survey. Figure 1 illustrates this sampling procedure. In the diagram each of the 39 'schools' is represented by a segment delineated by short ticks. The urbanicity-by-minority categories are delineated by longer ticks, and the points drawn by systematic sample of size 10 are indicated by long dotted ticks. The segment of the corresponding school is underlined. The asterisks under the segments indicate the median income categories for the schools.

Arrangements for substitution of the non-cooperating schools in the sample are described in Koffler (1991).

2.1.3 Replicate groups – stratification

The design of the survey for a state has a number of features that cannot be explicitly modelled. The 'reference' model that is considered by the NAEP analysis staff as well as by other researchers is that

Table 1: Clustering structure of the New Jersey sample.

Groups	Numbers of selected students within clusters						
1 – 7	29+20	23+28	22+26	23+19	25+23	30+23	39+27
8 – 14	23+29	25+29	28	24+24	29+28	26+29	25
15 – 21	26+26	29+27	26	26+29	29+26	23+22	22
22 – 28	28+28	29+26	22+25	26+23	26+25+25	13+14	21+26
29 – 35	28+23	22+26	26+28	22+26	22+24	27+25	22+55
36 – 42	23+23	23+19	22+25	23+29	28+27	24+26	28+26
43 – 49	29+29	29+26	29+26	25+23	28+26	24+25	**54**+27
50 – 53	27+29	28+29	27+27	25+27+25			

Notes: Each entry of the table contains the numbers of selected students within the clusters in each group 1 – 53. For example, group 1 has a cluster with 29 and one with 20 students in the sample. Groups 26 and 53 have three clusters each in the sample. The count for the certainty school is printed in boldface.

of a stratified weighted clustered sampling. The 'strata' in this context are defined after selection of the schools. To avoid confusion with the strata defined by crossclassification of urbanicity, median income, and minority composition we refer to them as *replicate groups*. It was decided that each state would have 56 replicate groups (with a few exceptions). The procedure of forming replicate groups is described in Koffler (1991) and justified in Johnson and Rust (1992). Most of these groups comprise a pair of clusters, others have either three or just one cluster. For operational convenience, empty replicate groups (containing no clusters) are declared, so that the total number of groups is 56. The main purpose of this is to have a uniform format for the user tapes for all the states and territories.

2.1.4 Selecting students

The school districts were requested to compile lists of all the eighth-grade students in the selected schools. From each selected school with enrollment of more than 35 eighth-graders a random sample of 30 students was drawn without replacement. From schools with fewer than 36 students all students were included in the sample. In order to ensure that each student had approximately the same chance of being included in the sample, schools with fewer than 20 students were 'thinned' (preselected); they were excluded from the sampling frame with probability inversely proportional to the total number of students in small schools.

For example, the New Jersey sample is described by the clustering structure of selected students within schools, and of the selected schools within groups, given in Table 1. The sample contains one certainty school (in stratum 49) with 60 selected students of whom six did not cooperate with the survey. A few small schools were aggregated into clusters containing at least 20 students each. Most clusters have between 20 and 30 students in the sample. The sample comprises 2710 students from 104 clusters (consolidated schools) in 53 groups; two groups are represented by three clusters each and four groups by one cluster each. In the process of selecting the sample of schools three small schools were 'thinned' out. There are three empty groups (54 through 56).

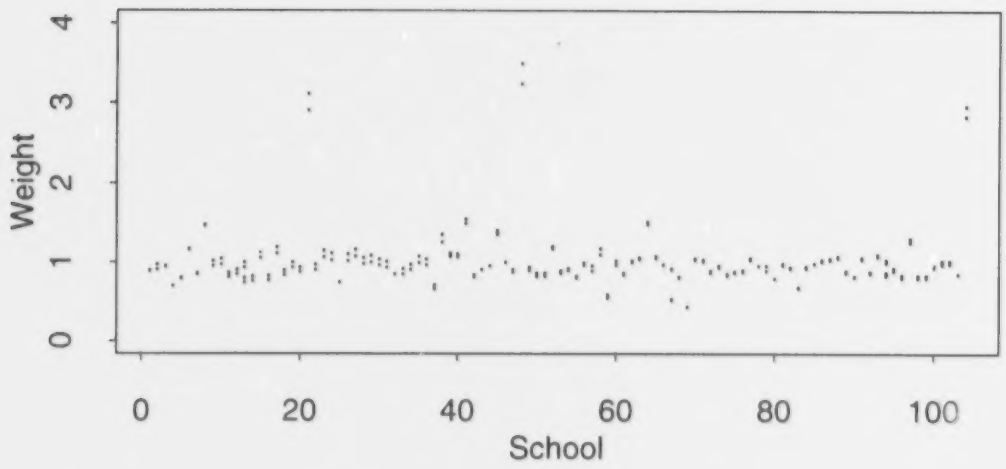

Figure 2: Adjusted weights for New Jersey.
Notes: The horizontal axis is the cluster index (1, ..., 104) and the vertical axis is the adjusted sampling weight. Duplicate values of the weight in a cluster are represented by a single dot.

2.1.5 Sampling weights

Each school (not cluster) is associated with a *school-level* design (base) weight and each student with a *student-level* (within-school) design weight. The sampling of students is conditionally independent of the sampling of schools, given the selected set of schools, and so each student's design sampling weight is equal to the product of these two weights. After the sampling procedure and administration of the survey questionnaire the students' weights are adjusted for non-response. Figure 2 contains a plot of these *adjusted* weights for the New Jersey sample. Note that the weights have little variation within clusters; in fact, most clusters have only two distinct and in most cases not very different values of the weight. There are three clusters with weights about three times larger than the rest of the clusters.

2.2 Proficiency scores

The student questionnaire contains items of four kinds:

- socio-demographic and background
- attitudinal
- experiential
- cognitive.

The background items probe the family environment and the educational level of parents. Attitudinal items relate to student's familiarity with calculators and computers, to perception of usefulness of mathematics,

and the like. Questions about mathematics classes taken are an example of experiential items. Cognitive items are mostly multiple choice items, and are scored as correct or incorrect. Based on these scores a number of (sub-)scales are defined (Measurement, Data analysis and statistics, Geometry, and the like). Our discussion is restricted to the *composite scale*, based on all the cognitive items.

For a given scale a proficiency score is defined for each student. It is estimated from the item-level scores by an item-response method (see Lord, 1980, and Mislevy, 1985, for background). The estimation is strengthened by incorporating (conditioning on) information contained in the student- and school-level background variables. The proficiency scores are subject to uncertainty, and they are represented for each student by a set of five *plausible values*.

The teacher questionnaire contains items about the teacher's qualifications, teaching methods, and about emphasis on elements of the curriculum.

Figure 3 presents a compact graphical summary of the proficiency scores and final weights. The proficiency scores are represented by the first set of plausible values. The plots on the left-hand side summarize the distribution of proficiency; at the top its values are plotted, and the within-cluster means are joined by a solid line; at the bottom the within-cluster standard deviations are plotted. The right-hand plots summarize the association of sampling weights and proficiency. At the top the two sets of quantities and at the bottom their within-cluster means are plotted.

2.3 Notation

In general, we use capitals to denote quantities that refer to the population, and lowercase characters to denote sample quantities. For example, N stands for the population size (number of students in the population), and n for the sample size (number of students in the sample). The proficiency score for student i (in the population) is denoted by Y_i. For students in the sample we use three indices, ijk, for student $i = 1, \ldots, n_{jk}$ in cluster $j = 1, \ldots, m_k$ in group $k = 1, \ldots, K$, and by y_{ijk} we denote the proficiency score of student ijk. Implicitly, we have introduced n_{jk} as the number of sampled students in the cluster jk and m_k as the number of sampled clusters in group k. The population counterparts for n_{jk} and m_k are denoted by N_{jk} (number of students in cluster jk, $j = 1, \ldots, M_k$) and M_k, respectively. The proficiency scores and the adjusted (final) sampling weights for the sampled students are denoted by y_{ijk} and w_{ijk}, respectively. For the plausible values we use another index, $h = 1, \ldots, 5$, so that y_{ijkh} is the plausible value h for student ijk.

Model parameters are also denoted by lowercase, such as μ, and their estimators are denoted by the same characters with 'hats', such as $\hat{\mu}$. When there are several estimators for a single parameter they are distinguished by an (additional) index. In notation we do not distinguish between an estimator (a function of the data, considered as a random variable) and an estimate (the realized value of the estimator for the drawn sample). The sampling variance of an estimator, say $\hat{\mu}$, is denoted by $\mathrm{var}(\hat{\mu})$, and the estimate and estimator of this variance is denoted by $\widehat{\mathrm{var}}(\hat{\mu})$.

For random variables we use lowercase Greek characters $\alpha, \beta, \ldots, \varepsilon$, and for parameters μ (mean), σ^2 (variance), ρ (correlation), τ (variance ratio), and the like. The expectation of a random variable, say γ,

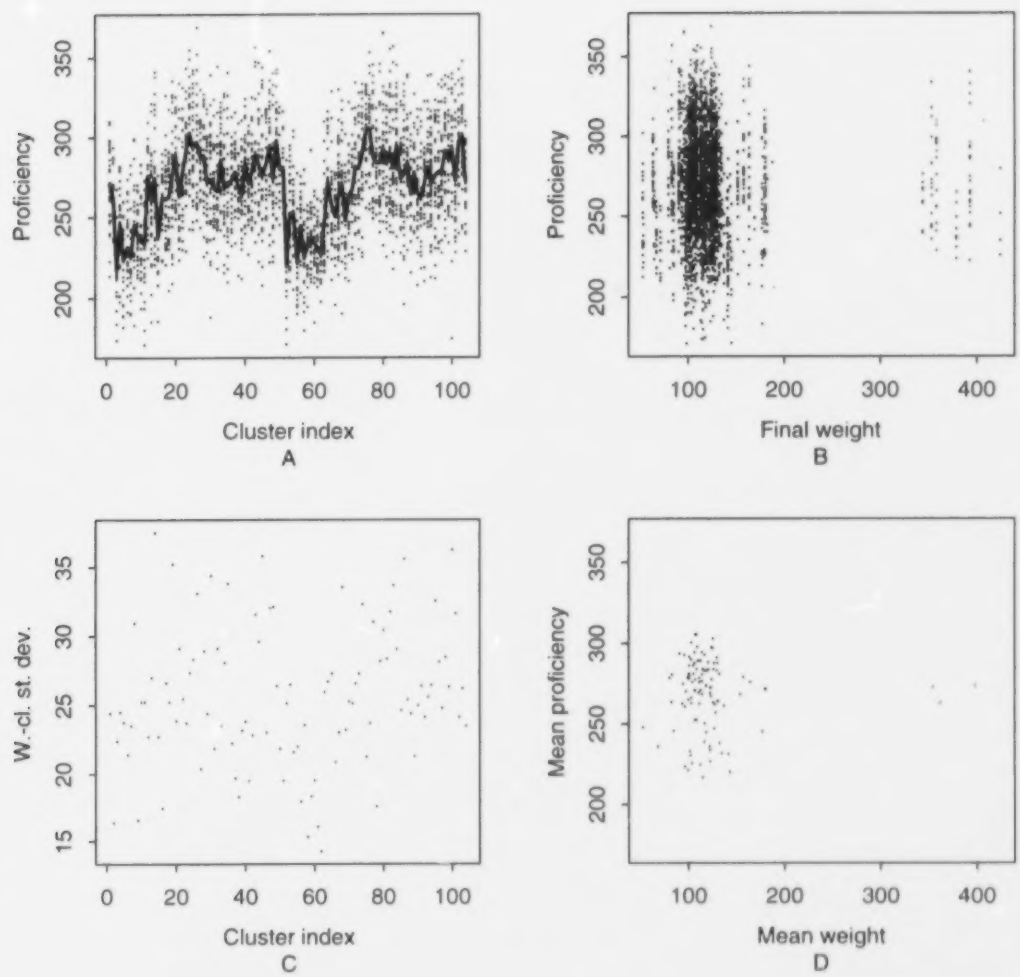

Figure 3: Final weights and proficiency scores for the New Jersey sample.

Notes: The plots are: A. Proficiency by cluster, with within-cluster mean proficiencies joined by solid line, B. Proficiency by final weight (students), C. Within-cluster standard deviations of proficiency, and D. Within-cluster mean proficiency by mean weight.

is denoted by $\mathbf{E}(\gamma)$. When it is necessary to distinguish between taking expectation over the samples and over the students this is explicitly stated.

Vectors and matrices are denoted by bold characters, Latin for constants, Greek for random vectors and matrices.

3 Model-based estimation of standard errors in stratified clustered sampling

In this section the model-based method of Potthoff, Woodbury, and Manton (1992) is implemented for estimation of the population and subpopulation means in the stratified clustered sampling design used in the NAEP State Assessment Program. The method relies on a superpopulation approach and has several features of the standard analysis of variance.

In the previous section we identified the following features of the sampling design:

1. sampling weights;
2. clustering (students within schools);
3. stratification (replicate groups);
4. non-response;
5. indirect measurement (estimation) of the outcome.

We consider first models and estimation procedures that accomodate each of these features on their own, and then construct a model that combines all of these features.

3.1 Modelling features of the design

3.1.1 Sampling weights

Let Y_i, $i = 1, 2, \ldots, N$, be the proficiency scores for the population of N students (say, in a state). The population mean is defined as

$$\mu = \frac{\sum_i Y_i}{N}$$

(the summation is over the entire population). When proficiency scores are available only for the sampled students, y_i, $i = 1, 2, \ldots, n$, the population mean is commonly estimated by the weighted mean

$$\mu = \frac{\sum_i w_i y_i}{\sum_i w_i} \tag{1}$$

(the summations are over the sampled students), where w_i is the sampling weight associated with student i. When the weights w_i are constant, equation (1) coincides with the unweighted mean,

$$\mu = \frac{\sum_i y_i}{n}.$$

In this case the common (non-zero) value of the weights is irrelevant for the estimator in (1). More generally, applying a (positive) constant multiplicative factor on the weights (that is, changing w_i to Cw_i for some $C > 0$) does not affect the estimator of the mean (1).

For equal weights the sampling variance of the estimator (1) is estimated by

$$\widehat{\text{var}}(\mu) = \frac{\sum_i (y_i - \mu)^2}{n - 1}.$$

A natural extension for unequal weights is the equation

$$\widehat{\text{var}}(\mu) = \frac{\sum_i w_i (y_i - \mu)^2}{\sum_i w_i - 1}. \tag{2}$$

Equation (2) is not invariant with respect to constant multiplicative factors of weights, and so a 'reasonable' choice of *normalization* for the weights w_i is essential. In one normalization the sample mean of the weights is equal to unity, that is, $\sum_i Cw_i = n$, or $C = n/\sum_i w_i$. Another choice is defined by the requirement that the total of the weights be equal to the total of the squares of the weights:

$$\sum_i Cw_i = \sum_i C^2 w_i^2,$$

that is, $C = \sum_i w_i / \sum_i w_i^2$. For a probability random sample the total of the weights normalized in this manner, $(\sum_i w_i)^2 / \sum_i w_i^2$, is referred to as the effective sample size (Potthoff *et al.*, 1992). For this normalization of weights the estimator in (2) is unbiased.

The sample size n is greater or equal to the effective sample size n_A. Their ratio, n/n_A, is referred to as the *design effect due to weights*. It is equal to unity when the weights are constant.

3.1.2 Clustering

Clustering is usually represented in statistical models by a non-negative correlation among the observations within a cluster, or by cluster-specific 'effects' – deviations of within-cluster means (or other summaries) from the corresponding population summary. These two model approaches are essentially identical, and are often used interchangeably. In the latter aproach we have a decomposition of the variance of the outcomes into its within- and between-cluster components. These we denote by σ_W^2 and σ_B^2, respectively. The variance ratio τ is defined as $\tau = \sigma_B^2/\sigma_W^2$. The within-cluster correlation is equal to $(\rho =)\ \sigma_B^2/(\sigma_W^2 + \sigma_B^2) = \tau/(1 + \tau)$.

It is intuitively appealing to assume that the outcomes of students within a classroom are positively correlated because students in a classroom share the same educational processes and experiences. The size of this correlation exacts an influence on the estimators of the population mean. This can best be illustrated by considering two extreme cases. If the outcomes for students within a school are perfectly correlated, that is, they are equal, then the outcomes for the students from a school are perfectly summarized by the data for any one of the students. If the sample comprises n students from m schools ($m < n$) there are only m 'essential' observations, not n. On the other hand, if the within-school correlation vanishes there are n essential observations. In intermediate cases (small positive correlation), it is reasonable to expect

that the sample at hand is as informative (its sample mean has as small a sampling variance) as a random sample of somewhat smaller size. The ratio of these sample sizes is referred to as the *design effect due to clustering*.

The sampling variance of the arithmetic mean $\mu = \sum_i y_i/n$ is equal to

$$\text{var}(\hat{\mu}) \;=\; \frac{\sigma_W^2}{n} \sum_j \left(1 + \frac{\tau}{n} \sum_j n_j^2 \right) ; \tag{3}$$

it is an increasing function of the correlation ρ, and of the variance σ_W^2. Note that there may be more efficient estimators of the population mean than the arithmetic mean $\hat{\mu} = \sum_{ij} y_{ij} / \sum_j n_j$, in particular, when the clusters have a wide range of sample sizes. In an alternative estimator the influence of an outcome (its weight) from a large cluster would be smaller than from a small cluster. Clearly, for an efficient estimator these 'weights' have to depend on the within-cluster correlation ρ.

Clusters may have unequal within-cluster variances. Then equal within-cluster covariances across the clusters do not correspond to equal variance ratios.

3.1.3 Stratification

Stratification is an important device for reduction of sampling variation of the estimates from sample surveys. It can be interpreted as a partitioning of the target population into an exhaustive set of non-overlapping subpopulations called strata, and carrying out a separate sample survey for each stratum. The population mean and other quantities of interest can be estimated by combining the corresponding estimates for the strata. The target population may exhibit substantial variation, but variation within each *stratum* may be much smaller. The parameters referring to a stratum can then be estimated with high precision, even if such estimates are based only on a fraction of the sample.

Clearly, a key to successful stratification is in identifying a small number of strata with distinct stratum means or, more generally, attributes and characteristics strongly associated with the variables of interest.

We adopt the approach of the NAEP operational analysis and regard the 56 replicate groups as the strata. In standard survey practice strata are defined for the target population (or the sampling frame) *prior to sampling*. To avoid confusion with the stratification of schools in NAEP, defined by median income, minority composition, and urbanicity, we use the term *group* for each of the 56 replicate groups.

3.1.4 Adjustment for non-response and poststratification

Having been selected into the survey, individual students or entire schools may refuse to cooperate. If it is feasible within the practical constraints, a non-cooperating school is replaced by a 'substitute' which matches the selected school as closely as possible on several attributes (for instance, on the stratifying variables and the enrollment). Characteristics of the non-cooperating students are not known, and therefore a scheme for their replacement by cooperating students is not feasible. Instead, the sampling weights are adjusted to take account of the 'missing' observations. In decisions about (approximate) sample size due

account is taken for the expected proportion of non-cooperating students, as well as for differences between estimated and actual within-school enrollments and for a number of contingencies.

An important consequence of non-cooperation (nonresponse) is that the weights, proportional to the reciprocals of the probabilities of being selected into the sample are not proportional to the reciprocals of the probabilities of inclusion in the sample. When nonresponse is informative, using these design weights would result in biased estimation. The design weights are therefore adjusted for nonresponse; for details of weight adjustment in NAEP Trial State Assessment, see Koffler (1991). Since the adjustment depends on the sample drawn, the weights are random variables (a different sample may result in a different weight adjustment even for a student included in both samples). Of course, a student responding in one survey might not respond in a hypothetical replicate of this survey; however, we have no information about the consistency of the pattern of nonresponse.

As a naive model for weight adjustment we consider an underlying mean weight for each student, averaged over all possible samples (or all those in which the student would be included in the sample). Since weights are invariant with respect to constant multiples and are proportional to reciprocals of probabilities it is advantageous to use the logarithm scale. For each subject i we posit the model

$$\log(w_{i,s}) \;=\; \log(\bar{w}_i) + \gamma_{i,s}\,, \tag{4}$$

where $w_{i,s}$ is the sampling weight assigned to student i when sample s is drawn, $\bar{w}_i$ is the geometric mean of the weights for student i, taken over all the samples s, and $\{\gamma_{i,s}\}_s$ is a random sample from a centered distribution. The variance var(γ_i), taken over the hypothetical samples s is a measure of how influential the nonresponse is. The clustered nature of the weight adjustment can be incorporated by a variance component model:

$$\log(w_{ijk,s}) \;=\; \log(\bar{w}_{ijk}) + \gamma^{(c)}_{jk,s} + \gamma^{(c)}_{ijk,s}\,, \tag{5}$$

where $\{\gamma^{(c)}_{jk,s}\}$ and $\{\gamma^{(c)}_{ijk,s}\}$ are two mutually independent random samples and $\bar{w}_{ijk}$ is the geometric mean of the weights for student ijk.

In surveys that are carried out on well-researched target populations it is advantageous to adjust the design (sampling) weights so as to bring them into accord with information about the target population external to the survey (various 'official' sources, censuses, and the like). This is referred to as poststratification. Poststratification is not applied in NAEP State Trial Assessment.

3.1.5 Estimation of proficiency

The proficiency scale is defined in relation to the cognitive items. The proficiency of a student is estimated using item-response models; see Koffler (1991) for details, and Mislevy and Bock (1982) and Lord (1980) for background. In order to adequately represent the variation of the estimators of proficiency for each student the proficiency is represented by a set of five draws from the estimated posterior distribution of the proficiency. Specifically, the item-response method used (Mislevy and Bock, 1982) yields an estimated

distribution of the underlying parameters, from which five random draws are obtained and a set of five *plausible values* is calculated based on the drawn values. Estimation of proficiency scores is improved by conditioning on several background variables. For details, see Johnson and Allen (1992, Chapter 11) and Mislevy and Sheehan (1991).

In general, estimation of any parameter is carried out for each plausible value (five analyses), and the mean of these estimates is adopted. Formally, let $\mathbf{y} = \{y_{ih}\}_{ih}$ be the $n \times 5$ matrix of plausible values for the entire sample, and μ_h, $h = 1, \ldots, 5$, be the estimator based on the hth set of plausible values. Then the mean $\mu = \sum_h \mu_h/5$ is the adopted estimator.

To emphasize dependence on data we write $\mu_h = \mu(\mathbf{y}_h)$ where $\mathbf{y}_h$ denotes column h of $\mathbf{y}$. Note that for estimators linear in the data, such as the weighted mean, μ is equal to the same estimator using the within-subject means of the plausible values, that is $\mu = \mu(\bar{\mathbf{y}})$, where $\bar{\mathbf{y}}$ is the vector of row-wise means of $\mathbf{y}$.

For estimation of the sampling variance the estimators of the mean of the sampling variances from the five analyses is supplemented by the variance of the estimates;

$$\widehat{\mathrm{var}}(\mu) \;=\; \mathbf{E}\{\mathrm{var}_s(\mu_h)\} + \mathrm{var}_h(\mu_h) \tag{6}$$

(the subscripts s and h indicate averaging over samples, and over the five sets of plausible values, respectively).

It is instructive to consider a plausible value y_{ih} as a sum of the overall (subpopulation) mean, μ, deviation of the student's proficiency y_i from the mean μ, $\delta_i = y_i - \mu$, and deviation of the plausible value from the proficiency, $\varepsilon_{ih} = y_{ih} - y_i$. Assuming that these two sets of deviations, $\{\delta_i\}$ and $\{\varepsilon_{ih}\}$, are independent, a desirable property of any procedure for generating plausible values, the proficiencies $\{y_i\}_i$ have smaller variation than the plausible values $\{y_{ih}\}_i$, for any h. The difference is the variance of the plausible values around the proficiency score:

$$\mathrm{var}(y_{ih}) \;=\; \mathrm{var}(\delta_i) + \mathrm{var}(\varepsilon_{ih}).$$

Note that for the within-student means of plausible values $\bar{y}_i$ we have

$$\mathrm{var}(\bar{y}_i) \;=\; \mathrm{var}(\delta_i) + \sum_i \mathrm{var}(\varepsilon_i)/5.$$

These means exhibit more variation than the proficiency scores; the variance of the latter is $\mathrm{var}(\delta_i)$. It is therefore not appropriate to carry out a single analysis using the student-wise means $\bar{y}_i$.

3.2 Randomness and conditioning in inference

In surveys, as in statistical practice in general, we are interested in *sampling distributions* of estimators. An orthodox view of the sampling distribution of an estimator in a survey is to consider the distribution of the estimates of a parameter in a large (infinite) number of (hypothetical) replications of the survey. The goal in a typical estimation problem is to make inference about such a distribution, based on a *single realization*

of the survey. Clearly, features of the survey have to be utilized to compensate for lack of replication. For a survey with a complex sampling design, non-response, and imperfect reliability of the response (for instance, due to measurement or estimation error), a hypothetical replication will have different students and schools in its sample, it may have a different sample size (different numbers of students and schools), but even the response/outcome of the student who happens to be selected in both surveys will be different (students', or indeed, our responses to even the most ubiquitous survey items are known not to be perfectly reliable).

Unconditional inference, averaging over a large number of hypothetical surveys, is often a tall order, and in practice inference is conditioned on the selected sample. It is meaningful, we believe, to consider conditioning on the sample that would have been obtained had each selected school and individual fully cooperated. Such a conditional inference is difficult to conceptualize because selection of the students is conditional on selection of their schools, (some) schools that refused to cooperate were substituted by other schools in the survey, and so on. Moreover, a school that failed to cooperate in the realized survey might cooperate in a hypothetical replication of the survey.

3.3 Jackknife

This section describes the jackknife method used for estimation of population and subpopulation means and their standard errors. The jackknife is a general method for reduction of bias of estimators and for estimation of their sampling variances. We describe the jackknife method as applied to NAEP State Trial Assessment.

The mean proficiency for the state is estimated by the weighted mean

$$\mu = \frac{\sum_{ijk} w_{ijk} y_{ijk}}{\sum_{ijk} w_{ijk}}. \tag{7}$$

Computation of the sampling variance of this estimator presents problems arising from complexity of the sampling design: unequal probabilities of selection, clustered sampling design and adjustment for nonresponse.

Each of the $K = 56$ replicate groups (whether empty or not) is associated with a *pseudoanalysis* carried out on a *pseudosample*. In groups with more than one cluster the clusters are assigned order (first, second, etc.) at random. If group k contains two clusters then the pseudosample for pseudoanalysis k is created by replacing the first cluster in group k by the other cluster in the group. This is equivalent to doubling the weights for all students in the second cluster. For groups with three clusters the first cluster is removed, and the weights for the students in the other two clusters are multiplied by 1.5. Thus, each student in the sample is associated with $K + 1 = 57$ *replicate weights*. These weights are given in the NAEP dataset. If poststratification were applied the replicate weights would have to be adjusted by poststratification of the pseudosample. When carried out operationally, this represents a substantial computational load.

The kth pseudoanalysis evaluates the estimator (7) using the kth set of replicate weights; we denote this estimator by $\mu^{(k)}$. The jackknife estimator of the mean μ is defined as the arithmetic mean of the

pseudoestimators $\mu^{(k)}$:

$$\mu_J = \frac{\sum \mu^{(k)}}{K}. \tag{8}$$

The sampling variance of μ_J is estimated as the sum of squares of deviations of the pseudoestimators $\mu^{(k)}$ from the jackknife estimator μ_J:

$$\widehat{\text{var}}(\mu_J) = \sum_k (\mu^{(k)} - \mu_J)^2; \tag{9}$$

see Wolter (1985) for details. Note that for estimation of the population mean only groups with two or more clusters contribute to the sum of squares in (9). However, for subpopulation means the same ordering of clusters is used, and if the subpopulation is represented only by one non-empty cluster in the dataset which happens not to be the first cluster, the group does make a contribution to the sum of squares in (9). In practice the estimator (7) is used instead of (8), and (9) is used as the estimator of sampling variance of (7). In brief, the jackknife is used only for sampling variance estimation.

The jackknife method given by (8) and (9) appears to be easy to implement, although the size of the dataset is substantially inflated by the replicate weights. In NAEP Trial State Assessment the students' records have length of about 1700, but two sets of replicate weights take up more than 800 columns.

Note that instead of the weighted mean (7) other statistics or estimators can be used as the 'parent' method for the jackknife. Ordinary regression is an important example.

Our study focuses on methods of estimation of the population mean, and of the sampling variances of these estimators, that depend on the design only through a single set of weights, the clustering, and the stratification.

3.4 Model-based methods

This section describes the model-based approach of Potthoff *et al.* (1992), as applicable to the NAEP Trial State Assessment. In general, an estimator for the quantity of interest (say, the ratio estimator for the population mean) is considered, and its sampling variance is expressed as a function of the modelled features of the design. Typically, these features include clustering and stratification. Clustering can be represented by one or several variance components and stratification by stratum-specific means (parameters).

For the NAEP State Assessment we consider the superpopulation model

$$y_{ijk} = \mu_k + \delta_{jk} + \varepsilon_{ijk}, \tag{10}$$

where the group means $\{\mu_k\}$ are unknown constants and δ_{jk} and ε_{ijk} are mutually independent random variables with zero expectations and respective variances σ_B^2 and $\sigma_{W,jk}^2$. The within-cluster variances $\sigma_{W,jk}^2$ are positive and unknown, and the between-cluster variance σ_B^2 is a non-negative constant. Note that σ_B^2 is the covariance of two observations in the same cluster:

$$\text{cov}(y_{ijk}, y_{i'jk}) = \sigma_B^2 \qquad (i \neq i'). \tag{11}$$

Often a common within-cluster variance is considered, $\sigma^2_{W,jk} \equiv \sigma^2_W$. This is not a realistic assumption for NAEP State Trial Assessment, however.

We consider the weighted mean, or the ratio estimator,

$$\mu = \frac{\sum_{ijk} w_{ijk} y_{ijk}}{\sum_{ijk} w_{ijk}}. \tag{12}$$

Assuming (10) and appropriateness of the weights, that is, they are proportional to the reciprocals of sampling probabilities, μ is an unbiased estimator of the superpopulation mean

$$m = \frac{\sum_k W_k \mu_k}{W}, \tag{13}$$

where $W_k = \sum_{ij} w_{ijk}$ and $W = \sum_k W_k$. We denote $W_{jk} = \sum_i w_{ijk}$, $\mu_{jk} = W_{jk}^{-1} \sum_i w_{ijk} y_{ijk}$, and $\mu_k = W_k^{-1} \sum_{ij} w_{ijk} y_{ijk}$. Note that μ_{jk} and μ_k are unbiased estimators of μ_k.

3.4.1 The sampling variance of the mean

Following Potthoff *et al.* (1992) we consider the weighted means μ_{jk} as *aggregate* observations. We use the 'effective sample size' normalization of the weights; we set

$$w_{A,ijk} = \frac{W_{jk}}{\sum_i w^2_{ijk}} w_{ijk}, \tag{14}$$

so that

$$\sum_i w_{A,ijk} = \sum_i w^2_{A,ijk} = \frac{W^2_{jk}}{\sum_i w^2_{ijk}}, \tag{15}$$

and denote this total of weights by $n_{A,jk}$. In a general context, *Potthoff et al.* (1992) refer to $n_{A,jk}$ as the 'effective sample size', to emphasize a connection with the number of 'degrees of freedom' of certain variance estimators, see Section 3.4.2. For NAEP, $n_{A,jk}$ can be interpreted as the effective sample size of cluster jk, although these quantities cannot be compared across clusters. A counterintuitive example arises when there is a cluster with a large number of students, each with very small weight, and another cluster with a small number of students, each with very small weight. For non-empty clusters $1 \le n_{A,jk} \le n_{jk}$, and $n_{A,jk}$ approaches these extremes when the cluster contains a single observation with dominant weight ($n_{A,jk} \doteq 1$), and when all the weights are almost constant ($n_{A,jk} \doteq n_{jk}$). The latter is the case in the 1990 Math Trial State Assessment.

The within-cluster weighted mean is

$$\mu_{jk} = \frac{\sum_i w_{A,ijk} y_{ijk}}{n_{A,jk}}, \tag{16}$$

and its variance is

$$\mathrm{var}(\mu_{jk}) = \frac{1}{n^2_{A,jk}} \left(n^2_{A,jk} \sigma^2_B + \sum_i w^2_{A,ijk} \sigma^2_{W,jk} \right) = \sigma^2_B + \frac{\sigma^2_{W,jk}}{n_{A,jk}}. \tag{17}$$

The statistics μ_{jk} are mutually independent linear components of the estimator (12);

$$\mu = \frac{\sum_{jk} W_{jk}\mu_{jk}}{W}.$$

We define the effective cluster sample size as

$$n_B = \frac{W^2}{\sum_{jk} W_{jk}^2}$$

and the normalized weights as

$$w_{B,jk} = \frac{W}{\sum_{jk} W_{jk}^2} W_{jk}.$$

Now

$$\mu = \frac{\sum_{jk} w_{B,jk}\mu_{jk}}{n_B},$$

and the sampling variance of μ is

$$\begin{aligned} \mathrm{var}(\mu) &= \frac{1}{n_B^2}\sum_{jk} w_{B,jk}^2 \left(\sigma_B^2 + \frac{\sigma_{W,jk}^2}{n_{A,jk}}\right) \\ &= \frac{\sigma_B^2}{n_B} + \frac{1}{n_B}\frac{\sum_{jk}\sigma_{W,jk}^2 \sum_i w_{ijk}^2}{\sum_{jk} W_{jk}^2}. \end{aligned} \tag{18}$$

Thus the sampling variance of μ depends on the variance components σ_B^2 and $\sigma_{W,jk}^2$. The next section deals with estimation of these variances. Alternatives, discussed later, include imputing values for these variances and applying smoothing techniques to improve estimation of $\mathrm{var}(\mu)$ by pooling information across subsamples.

3.4.2 Estimation of the variance components

There are $\sum_{k=1}^{K} m_k + 1$ unknown variance parameters $\{\sigma_{W,jk}^2\}$ and σ_B^2 in (18). The within-cluster variances $\sigma_{W,jk}^2$ can be estimated as the weighted within-cluster corrected sums of squares

$$v_{A,jk} = \frac{1}{n_{A,jk} - 1}\sum_i w_{A,ijk}(y_{ijk} - \mu_{jk})^2. \tag{19}$$

They are unbiased estimators of $\sigma_{W,jk}^2$:

$$\begin{aligned} \mathrm{E}(v_{A,jk}) &= \frac{1}{n_{A,jk} - 1}\left\{\sum_i w_{A,ijk}\mathrm{var}(y_{ijk}) - n_{A,jk}\mathrm{var}(\mu_{jk}|\mu_{jk})\right\} \\ &= \sigma_{W,jk}^2. \end{aligned} \tag{20}$$

If a common within-cluster variance σ_W^2 is assumed the weighted sums of squares $\{v_{jk}\}$ can be pooled:

$$\sigma_W^2 = \frac{\sum_{jk}(n_{A\,jk}-1)v_{A\,jk}}{\sum_{jk} n_{A\,jk} - 1} \tag{21}$$

is an unbiased estimator of the common within-cluster variance σ_W^2. The definition of the effective sample sizes $n_{A\,jk}$ is motivated by unbiasedness of the estimators in (20) and (21). Also, these estimators have *approximate* distributions χ^2 with the degrees of freedom given by the denominators, see Potthoff *et al.* (1992).

For estimation of the between-cluster variance σ_B^2 we consider weighted within-group sums of squares;

$$v_B = \sum_{jk} u_{jk}(\mu_{jk} - \mu_k)^2, \tag{22}$$

where $\{u_{jk}\}$ is a suitable set of non-negative coefficients (weights). The expectation of v_B is

$$\begin{aligned} \mathbf{E}(v_B) &= \sum_{jk} u_{jk}\left\{\mathrm{var}(\mu_{jk}) + \mathrm{var}(\mu_k) - 2\mathrm{cov}(\mu_{jk},\mu_k)\right\} \\ &= \sum_{jk} u_{jk} U_{jk} \mathrm{var}(\mu_{jk}), \end{aligned} \tag{23}$$

where

$$U_{jk} = 1 - \frac{2W_{jk}}{W_k} + \frac{\sum_j W_{jk}^2}{W_k^2}.$$

The development thus far imposes no restriction on the coefficients u_{jk}. Obvious choices for them are the totals of sampling weights, W_{jk} and the effective sample sizes, $n_{A\,jk}$.

For $u_{jk} = W_{jk}$, (23) simplifies to

$$\mathbf{E}(v_B) = \sum_{jk}\left(W_{jk} - \frac{W_{jk}^2}{W_k}\right)\left(\sigma_B^2 + \frac{\sigma_{W\,jk}^2}{n_{A\,jk}}\right). \tag{24}$$

A drawback of the scheme based on $u_{jk} = n_{A\,jk}$ is that the effective within-cluster sample sizes $n_{A\,jk}$ cannot be compared across clusters, and may be very misleading when the sampling weights have a large between-cluster component of variation. Neither of these choices for $\{u_{jk}\}$ takes account of the differing within-cluster variances $\sigma_{W\,jk}^2$ or of the differential contributions of the clusters to the within-group sum of squares v_B in (22).

The within-cluster weighted means μ_{jk} have an approximate normal distribution, and so the squared deviation $(\mu_{jk} - \mu_k)^2$ has a χ^2-like distribution. Thus $\mathrm{var}\{(\mu_{jk} - \mu_k)^2\}$ is approximately proportional to the square of the expectation. The optimal choice of u_{jk} is given by the set of coefficients for which the variance of v_B is minimized (subject to a constraint, such as $\sum_k \sum_j u_{jk}$ is equal to a constant). Assuming, for the moment, that σ_B^2 is known, and ignoring the interdependence of the squared deviations, we obtain, using (17), the optimal coefficients

$$u_{jk}^* = \frac{1}{U_{jk}(\sigma_B^2 + \sigma_{W\,jk}^2 n_{A\,jk})}. \tag{25}$$

The within-cluster variances $\sigma^2_{W,jk}$ can be replaced by their estimates in (19). In the absence of an estimate of σ^2_B a guess has to be used. It turns out that the accuracy of this guess is not critical; for instance, setting $\sigma^2_B = 0$ in (25) is often adequate.

As an alternative, the reciprocals of the contributions to (23) can be calculated using an estimate of σ^2_B obtained by one of the other methods. In principle, this recursive algorithm can be applied until convergence, but changes after the first iteration are unimportant.

The between-cluster variance σ^2_B is estimated by the method of moments:

$$\sigma^2_B = \frac{v_B - \sum_{jk} u_{jk} U_{jk} \sigma^2_{W,jk}/n_{A,jk}}{\sum_{jk} u_{jk} U_{jk}} \tag{26}$$

Note that $U_{jk} = 0$ holds only when stratum k is represented in the sample by a single cluster. For such a cluster and stratum $\mu_{jk} = \mu_k$ and $\mu_{jk} = \mu_k$, and so these clusters make no contribution to the sum-of-squares statistics v_2 or v_B.

In Math 1990 Trial State Assessment most groups contain two clusters. As an alternative to the estimator (22) the following class of statistics for estimation of the between-cluster variance can be used:

$$v_2 = \sum_{k, n_k \geq 2} u_k (\mu_{1k} - \mu_{2k})^2, \tag{27}$$

where the summation is over all groups with at least two clusters, and $\{u_k\}$ are suitable weights (constants). A group with a single cluster in the sample cannot contribute to estimation of the between-cluster variation, and so the only apparent loss is due to the groups with more than two clusters. The only advantage of (27) over v_B in (22) is in relative computational simplicity. For a group with two or more clusters we have

$$\mathbf{E}(\mu_{1k} - \mu_{2k})^2 = \sum_{j=1}^{2} \mathrm{var}(\mu_{jk}) = 2\sigma^2_B + \sum_{j=1}^{2} \frac{\sigma^2_{W,jk}}{n_{A,jk}}, \tag{28}$$

and so

$$\mathbf{E}(v_2) = \sum_{k, n_k \geq 2} u_k \left(2\sigma^2_B + \sum_{j=1}^{2} \frac{\sigma^2_{W,jk}}{n_{A,1k}} \right). \tag{29}$$

This, together with (21), yields a class of moment estimators of σ^2_B:

$$\sigma^2_B = \frac{v_2 - \sum_k u_k \sum_{j=1}^{2} \frac{\hat{\sigma}^2_{W,jk}}{n_{A,1k}}}{2 \sum_k u_k}, \tag{30}$$

where the summations for k are over groups with at least two clusters in the sample. In analogy with the schemes for v_B we consider the following choices for the coefficients u_k:

- the within-group total sampling weights W_k,
- the total of the effective sample sizes $n_{A,1k} + n_{A,2k}$,
- weights inversely proportional to the expected sum of squares in (28) under $\sigma^2_B = 0$.

Table 2: Coefficients for the within-group sums of squares v_B and v_2.

SSQ Method	v_B (C)	v_2 (B)
I	W_k	W_{jk}
II	$n_{A,1k} + n_{A,2k}$	$n_{A,jk}$
III	$1/(2\sigma_B^2 + \sum_1^2 \sigma_{W,jk}^2/n_{A,jk})$	$1/\{U_{jk}(\sigma_B^2 + \sigma_{W,jk}^2/n_{A,jk}\}$
R	as III, with σ_B^2 in place of σ_B^2	as III, with σ_B^2 in place of σ_B^2

Note: Estimators of the between-cluster variance σ_B^2 are referred to by the combination of the sum of squares (SSQ) used **B** or **C** and the method (choice of coefficients), I, II, III, or R.

Alternatively, the variance σ_B^2 can be estimated by (30) using one of these sets of weights, and then reestimated using the weights inversely proportional to (28). Of course, this recursive estimation scheme can be used until convergence is achieved. However, after one such iteration the change in σ_B^2 is usually unimportant. The motivation for these sets of coefficients is analogous to their counterparts for (26).

The choices for the coefficients u_{jk} in v_B and u_k in v_2 are summarized in Table 2. Examples of these estimators are given in Section 3.5. We refer to the estimators of σ_B^2 and to the estimators of the sampling variance of μ by symbols 'B' (based on (30)) or 'C' (based on (26)), and 'I' (coefficients W_k or W_{jk}), 'II' (coefficients $n_{A,jk}$ or $n_{A,1k} + n_{A,2k}$), 'III' (reciprocals of the expected contributions to v_2 or v_B assuming $\sigma_B^2 = 0$), and 'R' (reciprocals of the expected contributions calculated for σ_B^2 estimated by the method I). The estimates of the variances $\sigma_{W,jk}^2$ and σ_B^2 are substituted for their true values in the identity for the variance of μ, (18).

If we insist on interpretation of σ_B^2 as a variance then negative values of σ_B^2 are not admissible. If for each negative value of (30) the estimate of σ_B^2 is set to zero, as is often done in practice, the resulting estimator is biased, especially when the true value of the parameter σ_B^2 is close to zero. On the other hand, $\kappa = \sigma_B^2$ can be interpreted as within-cluster covariance, see (11), and then its negative values are admissible. The minimum within-cluster covariance κ that can be realized for a cluster of size N_{jk} is $-1/(N_{jk} - 1)$. Note, however, that the sample cluster size n_{jk} may be much smaller than the population cluster size N_{jk}; a negative estimate of the covariance κ may be realizable for the sample selected from the cluster, but not for the entire population of the cluster.

3.5 Examples

The jackknife and model-based methods for estimation of the mean are illustrated on a few examples using the data from New Jersey and Oklahoma. Adjustment of the weights due to nonresponse is ignored throughout the section, but it is explored in the next section.

Table 3: Jackknife analysis. Estimation of the population mean of proficiency scores.

New Jersey	Plausible values					Overall
	1	2	3	4	5	
Weighted mean	269.47	269.42	269.37	269.40	269.65	269.46
Jackknife mean	269.48	269.43	269.37	269.42	269.65	269.47
JK stand. error	1.04	1.07	1.03	1.06	1.04	1.05
Oklahoma						
Weighted mean	262.95	262.74	262.78	262.71	262.63	262.76
Jackknife mean	262.91	262.71	262.71	262.68	262.60	262.73
JK stand. error	1.23	1.27	1.21	1.24	1.22	1.24

3.5.1 Population mean

Estimation of the population mean by jackknife is summarized in Table 3 for New Jersey and Oklahoma. In essence, separate jackknife analyses are carried out for each plausible value, and the estimators based on the plausible values are combined into the 'Overall' estimator which takes account of variation due to estimation of the proficiency scores. The estimate of the population mean proficiency is the average of the estimates of the mean based on the five plausible values. The sampling variance of the estimator of the population mean is estimated using (6). The estimates for New Jersey are based on 2710 students from 104 clusters, those for Oklahoma on 2222 students from 108 clusters.

The differences between the weighted means and the jackknife estimates of the means are inconsequential; note however, that the differences for Oklahoma appear to be consistent, though negligible. For most purposes the statistics based on plausible values 1 – 5 are of no interest, and their summaries in the right-most column of Table 3 are used.

Model-based estimators of the standard error of the weighted mean are summarized in Table 4 for the four estimators based on each sum-of-squares statistic, v_2 and v_B. The estimators BI and CI require least computation, owing to simpler equations for estimation of σ_B^2, and the estimators BR and CR most (almost twice as much as BI and CI, respectively). For completeness, the second row of the table contains the pooled estimates of the common within-cluster variance σ_W^2.

The eight sets of estimators of the standard error are within a range of 0.04, but they differ from the jackknife estimate by about 0.15 (almost 15 per cent). Based on this analysis we cannot arbitrate whether such a difference is due to sampling variation of the estimator of the standard error, or whether the jackknife and model-based estimators have different biases (or indeed, whether the jackknife is unbiased and model-based estimators are not).

Table 5 summarizes model-based estimation of the population mean for Oklahoma. Contrasting the analysis for New Jersey the model-based estimates for Oklahoma are very close to the jackknife estimate of

Table 4: Model-based estimators of standard error of the weighted mean.

	New Jersey					
	Plausible values					Overall
	1	2	3	4	5	
Weighted mean	269.47	269.42	269.37	269.40	269.65	269.46
σ_W^2	679.69	665.84	665.89	678.80	667.75	671.59
			Method **BI**			
σ_B^2	99.36	104.17	97.83	100.40	98.26	100.01
Standard error	1.19	1.21	1.17	1.19	1.18	1.19
			BII			
σ_B^2	97.68	99.25	91.14	98.96	96.85	96.78
Standard error	1.18	1.18	1.14	1.18	1.17	1.17
			BIII			
σ_B^2	101.21	107.44	94.82	107.40	104.50	103.07
Standard error	1.20	1.22	1.16	1.22	1.21	1.20
			BR			
σ_B^2	96.85	99.29	90.20	97.59	96.20	96.02
Standard error	1.17	1.18	1.14	1.18	1.17	1.17
			CI			
σ_B^2	107.40	108.29	100.07	101.32	103.91	104.20
Standard error	1.22	1.22	1.18	1.19	1.21	1.21
			CII			
σ_B^2	104.61	104.46	95.56	99.71	101.35	101.14
Standard error	1.21	1.21	1.16	1.19	1.19	1.19
			CIII			
σ_B^2	103.04	102.55	94.06	99.63	100.30	99.92
Standard error	1.20	1.20	1.15	1.19	1.19	1.19
			CR			
σ_B^2	104.53	104.91	95.29	100.65	102.02	101.48
Standard error	1.21	1.21	1.16	1.19	1.20	1.19

Notes: Estimation of the population mean of proficiency scores for New Jersey. The methods are described in the text and in Table 2; $\hat{\sigma}_W^2$ is the pooled estimate of the within-cluster variance; $\hat{\sigma}_B^2$ is the estimate of the between-cluster variance. The estimates are given for each plausible value and for the proficiency score (column 'Overall').

Table 5: Model-based estimators of standard error of the weighted mean. Estimation of the population mean of proficiency scores for Oklahoma.

	Oklahoma					
	Plausible values					Overall
	1	2	3	4	5	
Weighted mean	262.95	262.74	262.78	262.71	262.63	262.76
σ^2_W	654.75	672.11	675.53	678.84	652.18	666.68
			Method **BI**			
σ^2_B	114.60	119.77	112.67	116.81	114.08	115.58
Standard error	1.22	1.25	1.22	1.24	1.22	1.23
			BII			
σ^2_B	118.12	123.49	116.43	121.29	117.10	119.28
Standard error	1.24	1.26	1.24	1.25	1.23	1.25
			BIII			
σ^2_B	103.71	111.82	101.29	113.47	104.88	107.03
Standard error	1.18	1.22	1.17	1.22	1.18	1.20
			BR			
σ^2_B	120.18	199.58	114.54	118.70	115.25	117.65
Standard error	1.25	1.25	1.23	1.24	1.23	1.24
			CI			
σ^2_B	121.29	128.09	118.91	120.82	119.20	121.66
Standard error	1.25	1.28	1.25	1.25	1.24	1.26
			CII			
σ^2_B	119.13	125.64	117.22	119.04	116.28	119.46
Standard error	1.24	1.27	1.24	1.24	1.23	1.25
			CIII			
σ^2_B	121.86	128.86	120.10	122.52	119.49	122.56
Standard error	1.26	1.27	1.25	1.26	1.24	1.26
			CR			
σ^2_B	124.14	125.58	118.95	120.73	118.50	121.58
Standard error	1.26	1.27	1.25	1.15	1.24	1.26

the standard error. The pooled estimates of the within-cluster variance for New Jersey and Oklahoma are alike, as are the between-cluster (within-group) variances. The latter variances are very large, considering purposeful grouping of the clusters into groups. For example, the estimated within-cluster correlations for Oklahoma, using method I, are around $115.58/(115.58 + 666.68) \doteq 0.15$. Without adjustment for group (stratification) these correlations are much larger (around 0.35).

The elapsed time for the analysis producing the eight model-based estimates displayed in Tables 4 and 5 is less than twice the elapsed time for the jackknife analysis.

Of principal interest in Tables 4 and 5 are the right-most columns ('Overall') giving estimates that take account of inaccuracy in estimation of the proficiency scores.

3.5.2 Subpopulation means

Table 6 displays the estimates of the means for a selected set of subpopulations in New Jersey. The subpopulation is characterized by the questionnaire item, and response; for instance, (183,2) in the first column of the table signifies the subpopulation of students who responded '2' (Graduated from high school) to item number 183 (Parent's educational level). For each subpopulation the corresponding sample size and number of (non-empty) clusters are given. The standard errors estimated by jackknife are given in parentheses underneath the associated estimate. For the model-based methods the standard errors are given in parentheses, and the estimated between-cluster variances in brackets. The methods B (for pairs of clusters) differ from their method C counterparts, but the differences are insubstantial in comparison with the estimators within a method, especially for small samples. To conserve space, only results for the method C are given.

There appears to be considerable agreement between the jackknife and model-based estimators of the standard errors, especially for larger datasets (with more than 1000 students). On the other hand, among the estimated standard errors for small datasets there are considerable differences. It is feasible, however, that they merely reflect substantial sampling variation. For instance, the dataset for item and response (28,2) (Asian American students), contains 131 students, 14 of whom are in a single cluster; of the remaining 56 non-empty clusters only 16 contain more than two students, and none contains more than six.

3.6 Adjustment of weights for nonresponse

For purposes of statistical analysis, adjustment is commonly interpreted as a perturbation of the sampling weights. The adjustments for a student in the sample drawn may be different from the adjustment in a different sample in which the student is also selected. This creates problems with all methods that rely on the sampling weights being constants fixed prior to selection of the sample. A simplistic approach to dealing with such adjustment is to ignore the stochastic nature of the adjusted weights (their variation over samples), and proceed with the analysis as if the adjustment of weights took place prior to sample selection. This approach is certainly justified when the weights are altered only marginally. This is the case in the New Jersey dataset but not in the dataset for Oklahoma. In this section we show, though,

Table 6: Jackknife and model-based estimates for a selected set of subpopulations; New Jersey.

New Jersey							
Item and response	Students and clusters	Weighted mean and σ^2_W	Jackknife	Method C			
				I	II	III	R
28,1	1789/95	279.19 662.56	279.54 (1.06)	(1.19) [67.50]	(1.14) [60.63]	(1.28) [84.29]	(1.33) [90.74]
28,2	398/61	240.79 397.64	240.83 (2.30)	(2.34) [73.73]	(2.12) [52.88]	(2.30) [76.45]	(3.20) [110.41]
28,4	131/57	296.97 655.03	296.99 (4.92)	(5.04) [116.43]	(5.17) [90.92]	(5.23) [160.63]	(4.83) [17.96]
31,1	656/25	279.25 719.71	279.24 (2.36)	(2.20) [72.66]	(2.04) [58.76]	(1.99) [55.28]	(1.97) [53.68]
31,2	1170/45	274.79 690.37	274.84 (2.10)	(1.54) [63.62]	(1.57) [66.69]	(1.57) [67.39]	(1.58) [68.10]
183,2	633/102	258.87 574.89	258.89 (1.59)	(1.76) [138.10]	(1.71) [127.06]	(1.76) [139.69]	(1.56) [95.48]
183,4	1225/104	281.38 716.87	281.38 (1.35)	(1.40) [96.56]	(1.36) [89.30]	(1.42) [100.87]	(1.41) [98.19]
193,1	386/34	281.73 515.23	281.70 (4.13)	(4.48) [364.58]	(3.83) [277.57]	(4.19) [322.51]	(4.02) [306.86]
193,2	677/64	272.97 560.91	273.00 (2.55)	(2.11) [150.46]	(2.05) [140.76]	(2.49) [221.61]	(2.60) [223.46]
193,3	1346/91	265.80 580.20	265.83 (1.97)	(1.89) [193.71]	(1.85) [183.92]	(1.98) [212.05]	(2.06) [233.36]
193,9	301/39	261.36 523.84	261.35 (4.61)	(3.88) [79.09]	(3.72) [36.80]	(4.59) [127.87]	(5.29) [103.27]

Notes: For each method the estimates of σ^2_B are given in brackets, [], and the estimated standard errors in parentheses ( ). The items and response options are: 28 – Derived race/ethnicity (1 – White, 2 – Black, 4 – Asian); 31 – Minority stratum;
183 – Parents' educational level (2 – Graduated from high school, 4 – Graduated from college);
193 – Teacher's graduate major (1 – Mathematics, 2 – Education, 3 – Other, 9 – Missing).

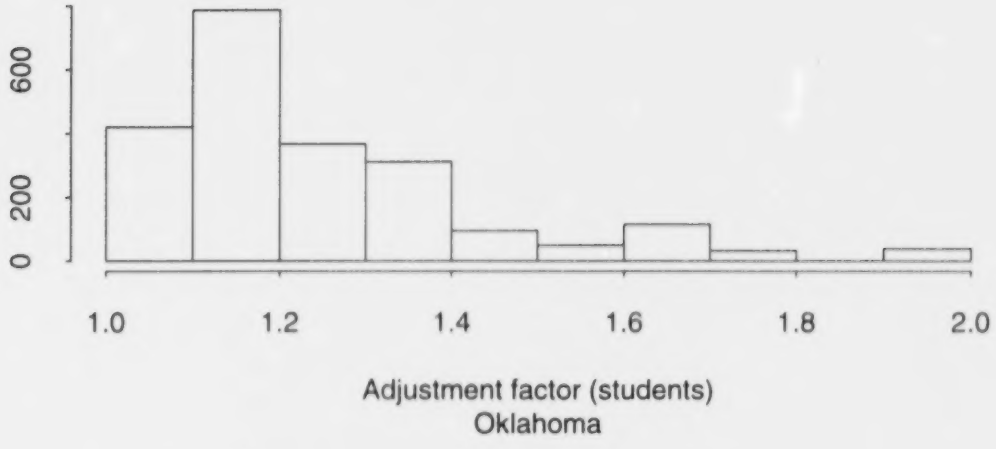

Figure 4: Student-level weight adjustment factors for Oklahoma.

that this approach involves trivial imprecision even for Oklahoma, so long as the adjustment of weights is worthwhile.

The design weights have two multiplicative components, school- and student-level design weights. The school-level weights are proportional to the reciprocals of the probabilities of inclusion of the school in the sample, assuming that all schools would cooperate. Note that these weights refer to schools, not to clusters. The student design weights are proportional to the conditional probabilities of inclusion in the sample, given that the school is included.

Adjustment of these weights due to non-response also has two multiplicative components – one for *clusters* and one for students. In the New Jersey dataset there are only three different cluster-level adjustment factors: 1.000 (no adjustment), 1.027, and 1.046, for 46, 37, and 21 clusters, respectively. The student-level adjustment factors are in the range 1.016 – 1.150, with mean and median equal to 1.06, and sample standard deviation equal to 0.024. In summary, the adjustments of the weights (product of the school- and student-level adjustments) are in the range 1.016 – 1.171, with mean and median equal to 1.08 and standard deviation equal to 0.032.

In contrast, in the dataset for Oklahoma (108 schools with 2222 students), where the non-response was much higher (about 20 per cent at student level), the adjustment of weights is much more substantial. The design weights (both school- and student-level) are constant within schools, and so are the school-level adjustments. The student-level adjustment factors have 35 distinct values, two in most clusters. These factors are in the range 1.00 – 1.95, see Figure 4.

The school-level adjustment factors are much less important. Essentially, there are three distinct values of the factor: for 60 schools (1195 students) the factor is equal to 1.00; for 42 schools with 895 students the adjustment factor is 1.015 – 1.016; and for the remaining 6 schools (132 students) the adjustment factor is 1.16. New Jersey and Oklahoma represent two extremes among the states participating in the 1990 Math

Trial State Assessment in terms of nonresponse and consequent adjustment.

Approaches other than jackknife to inference from survey samples regard the adjusted weights as design weights, ignoring the stochastic nature of the adjustments. We explore whether such an approach is justified for the model-based estimators by a Monte Carlo study in which the sampling weights are perturbed by random terms with suitably chosen dispersion. Instead of the realized weights w_{ijk} we consider a set of 'perturbed' weights, $w_{ijk}^{(p)}$, generated by the model

$$\log(w_{ijk}^{(s)}) \;=\; \log(w_{ijk}^{*}) + \delta_{jk}^{(w)} + \varepsilon_{ijk}^{(w)}, \tag{31}$$

where $\{\delta_{jk}^{(w)}\}$ and $\{\varepsilon_{ijk}^{(w)}\}$ are two mutually independent random samples from $\mathcal{N}(0, \sigma_{w,B}^2)$ and $\mathcal{N}(0, \sigma_{w,W}^2)$, respectively, and w_{ijk}^{*} is the mean weight for student ijk, averaged over all the samples. The choice of values for the variances $\sigma_{w,B}^2$ and $\sigma_{w,W}^2$ is discussed below. Note that $\exp(\delta_{jk}^{(w)})$ and $\exp(\varepsilon_{ijk}^{(w)})$ are essentially different from the weight adjustments; if they were not, weight adjustment would have no stochastic component. The model in (31) assumes independent deviation factors for clusters and students, both log-normally distributed. This assumption cannot be checked because the 'true' weights $w_{ijk}^{(s)}$ are not known. Also, the weights $\{w_{ijk}^{*}\}$ are defined subject to a multiplicative factor constant for a sample. In (31) a suitable constant factor is assumed. There is no evidence of dependence of the applied school- and student-level adjustment factors.

We adopt the 'working' assumption that the sampling weights are unbiased estimators of a fixed multiple of the reciprocal of the sampling probabilities, and that the adjustment of the design weights at both cluster and student levels has the following properties. Each adjustment factor has two components: adjustment of bias of the design weights, and a random component. We assume that the variation of the random component is of the same order as the variance of the bias of adjustment, that is, the adjustment is reasonably efficient. This suggests the choice of school- and student-level variances $\sigma_{w,B}^2$ and $\sigma_{w,W}^2$ of the same order of magnitude as the sample variances of the logarithms of the adjustment factors.

The adjustment factors for New Jersey are so small, that even for an unrealistically large perturbation of weights the weighted means have observed variances negligible in comparison with the estimated sampling variance of the estimator for the population mean. We chose the variances $\sigma_2^2 = 0.05^2$ for schools and $\sigma_1^2 = 0.1^2$ for students. To illustrate this perturbation, the basic descriptive statistics of the adjustment factors are compared with a random sample from the distribution used for perturbing the sampling weights. The basic descriptive statistics (minimum, median, mean, maximum, and standard deviation in parentheses) for the school-level adjustment (on log-scale) are:

0.000, 0.018, 0.027, 0.045, (0.018)

The same statistics for a random sample of size 104 (the number of schools) from the distribution generating the perturbation factors is

0.002, 0.044, 0.037, 0.157, (0.034)

The corresponding statistics for the student-level adjustments and perturbation are

0.016, 0.058, 0.056, 0.140, (0.022)

and

$$0.000, \quad 0.080, \quad 0.068, \quad 0.394, \quad (0.060)$$

Thus the simulated perturbation changes the weights much more than the realized adjustment for nonresponse.

One hundred sets of perturbed weights were generated for the entire sample and several subsamples of the New Jersey dataset. The mean and standard deviation of the simulated estimates of the population mean are 269.43 and 0.2035, respectively. The jackknife and ratio estimates of the population mean are 269.47 and 269.46, respectively. For subpopulations the corresponding differences are also trivial. For example, the mean of the simulated estimates of the mean proficiency of the Asian American students (131 students in 57 clusters) is 296.91 (standard deviation of the simulated estimates is 0.47), while the ratio and jackknife estimates are 296.99 and 296.97, respectively. The estimated standard error of these estimators is aroud 5.0. The variation in the estimates of the sampling variation is also unimportant.

A similar analysis for Oklahoma yields somewhat larger differences. The jackknife and ratio estimates of the population mean are 262.91 and 262.95, respectively, and the mean of the simulated estimates is 262.77 (the standard deviation of these estimates is 0.30). The corresponding means for the Asian American students (36 students in 25 clusters) are 286.47 (ratio estimate), 286.57 (jackknife), and 286.49 (simulation). The differences among these means are trivial in comparison with the substantial sampling error. The impact of perturbation of weights on the estimated sampling variance is also trivial.

3.7 Association of weight adjustment and proficiency

A simpler, though incomplete, way of assessing the influence of the weight adjustment on the estimate of (sub-)population means is based on exploring the association of the weight adjustment with the proficiency. For simplicity we consider the first plausible value as a representation of the proficiency. The estimate of the population mean for New Jersey, based on the design weights, is 269.28, 0.21 lower than the ratio estimate based on the adjusted weights. For Oklahoma, the design-weight sample mean is 262.95, 0.26 lower than the adjusted-weight sample mean. Such differences are no ionger trivial although the biases incurred are in no way consequential.

Influence of the weight adjustment on the estimate of the population mean is a resul of association of the school- and student-level adjustment factors with proficiency. Figure 5 displays the plot of the school-level adjustment of the weights against the school-means of proficiencies (left-hand panel) and the plot of student-level adjustment against the proficiencies. The school-level adjustments are positively associated with mean proficiency – 'better' schools were more likely to decline participation in the survey. On the other hand, student-level adjustment is negatively associated with proficiency. Students with lower ability are more likely to abstain from the survey. Since the school-level adjustment is on a much narrower scale, the overall adjustment is affected only very moderately by the school-level weight adjustments.

For subpopulations the impact of weight adjustment varies depending on the stochastic mechanism of 'selection' of the subpopulation. For example, the weighted means for urbanicity stratum 1 (442 students) in New Jersey are 238.01 and 238.14 for the design and adjusted weights, respectively; for students who

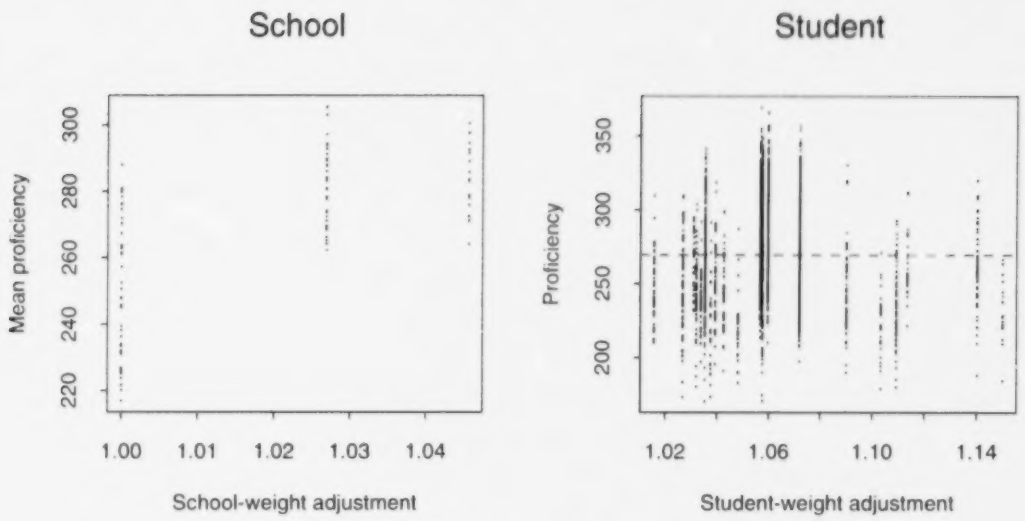

Figure 5: Association of cluster- and student-level weight adjustment factors with proficiency; New Jersey.

responded '3' (undecided) to the item about their perception of mathematics (555 students) these means are 259.80 and 259.63. Such differences are small fractions of the corresponding estimated standard deviations of the weighted means. The corresponding differences for Oklahoma are only marginally larger.

In conclusion, weight adjustments have a small, although perceptible, impact on estimates of the population and subpopulation means.

The adjustment of the sampling weights for Oklahoma is comparable to the perturbation of the school-level weights by $N(0, 0.1^2)$ and of the student-level weights by $N(0, 0.025^2)$.

3.8 Multivariate outcomes

It is easy to see that both the jackknife and the model-based methods have direct extensions for multivariate outcomes. Estimation of the population mean is carried out component-wise, and the sampling variance matrix of the vector of estimated means for the jackknife is

$$\widehat{\text{var}}(\mathbf{m}_J) = \sum_k (\mathbf{m}^{(k)} - \mathbf{m}_J)(\mathbf{m}^{(k)} - \mathbf{m}_J)^\top,$$

using the notation analogous to that in (9). For the model-based methods all the equations in Section 3.4 apply, with the variance components replaced by the corresponding variance matrices.

3.9 Modelling approach

In this section we consider an adaptation of the maximum likelihood method for estimation of the population mean. For future reference we consider ordinary regression instead of the population mean.

A well established approach to regression analysis of data from surveys with complex design relies on a superpopulation model in which survey features are typically represented as differences among sampling units. In the case of simple regression it is natural to consider regression of y on x with coefficients varying across the clusters, and differences (unknown constants or functions) among the groups:

$$y_{ijk} = a_{jk} + b_{jk}x_{ijk} + \varepsilon_{ijk}\,, \tag{32}$$

where $(a_{jk},\ b_{jk}) \sim \mathcal{N}\{(A_k, B_k),\ \boldsymbol{\Sigma}_k\}$ and $\varepsilon_{ijk} \sim \mathcal{N}(0, \sigma^2_{jk})$, independently. Usually, submodels of (32) defined by constraints, such as $\boldsymbol{\Sigma}_k \equiv \boldsymbol{\Sigma}$, $b_{jk} \equiv b_k$, and the like, are considered. Realistic submodels often still contain a large number of parameters, one for each group, so as to reflect substantial between-group differences. To consider the population mean, set $b_{jk} \equiv 0$ in (32).

For the case of constant weights, $w_{ijk} \equiv 1$, fitting such models, say, by maximum likelihood (ML), is carried out by iterative procedures the complexity of which depends essentially on the number of estimated parameters. For unequal weights the crossproduct statistics required for ML are replaced by their weighted versions. Interpretation of the estimates, as well as of the model parameters, is problematic because they have to be combined to obtain quantities that relate to the target population. A simple adaptation due to Harville (1974) adjusts for the bias of the maximum likelihood estimator of a variance due to ignoring the regression parameters (or the population mean). This method is called the restricted maximum likelihood (REML).

The principal disadvantage of the model in (32) is that no distinction is drawn between sampling errors and imperfect description of the association in the target population. On the other hand, the model-based procedures appear not to cater for separate components of variation stemming from the clustered nature of the target population. This deficiency can, in principle, be resolved by defining more complex population summaries such as measures of between-cluster variation.

4 Simulations

The purpose of the simulation study described in this section is to compare the properties of the jackknife and the proposed model-based estimators. The mean squared errors of the estimators of the (sub-)population means are of principal interest. In summary, the model-based estimators are much more efficient, in terms of mean squared errors, than jackknife, and the differences among the model-based estimators are relatively unimportant. We note, however, that the comparison is somewhat unfair to the jackknife since the data are simulated according to the model on which the alternative methods are based. In particular, weight adjustment is ignored in the simulations.

We consider a dataset, such as the set of all students in the New Jersey sample, with their sampling weights, clustering structure, and stratification/grouping, and replace the outcome variable by a set of values generated using the model in (10), with realistic values of the parameters $\{\sigma^2_{W\,jk}\}$, σ^2_B, and $\{\mu_k\}$.

Thus, the group-means $\{\mu_k\}$ are drawn independently from $\mathcal{N}(250, \Sigma^2_{str})$, and the within-cluster standard deviations $\{\sigma_{W\,jk}\}$ are drawn independently from $\mathcal{U}(V_L, V_H)$. The sets of group means and the

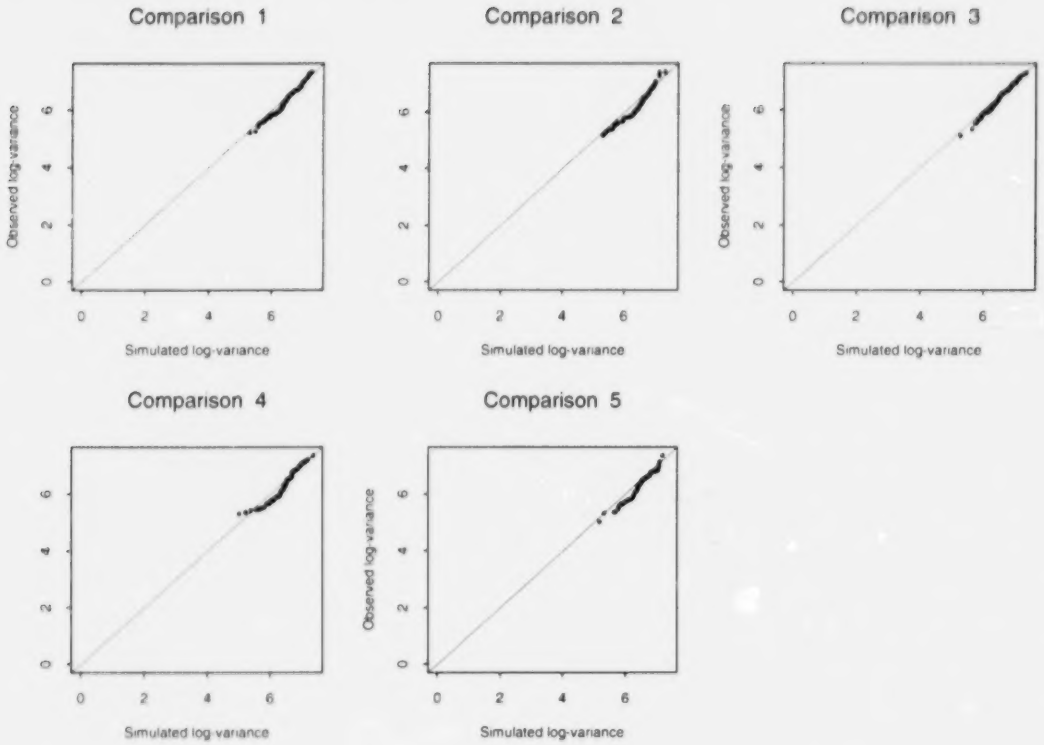

Figure 6: Comparison of the estimated and simulated within-cluster log-variances.
Notes: 'Comparison' k, $k = 1, \ldots, 5$, is the plot of the ordered values of the logarithms of the simulated within-cluster variances against the ordered values of the logarithms of the estimated within-cluster variances for the kth set of plausible values. The five simulations are mutually independent.

within-cluster variances are common to all datasets within a set of simulations, but different student- and cluster-level deviations are drawn independently for each simulated dataset.

For example, for the entire sample $\Sigma_{str} = 10$, $V_L = 10$, $V_H = 40$, and $\sigma_B^2 = 100$ generate datasets with features similar to those of the survey dataset. Ignoring the within-cluster variation of the sampling weights, the estimates of the within-cluster variances $\{\sigma_{W,jk}^2\}$ have the $\sigma_{W,jk}^2$-multiple of the χ^2 distribution with $n_{A,jk} - 1$ degrees of freedom.

The plots in Figure 6 compare the estimates $\{\sigma_{W,jk}^2\}$ for the five sets of plausible values for the New Jersey dataset with five mutually independent sets of simulated estimates of the within-cluster variances, drawn as realizations of the distributions $\sigma_{W,jk}^2 \times \chi_{n_{A,jk}-1}^2/(n_{A,jk} - 1)$, where the variances $\sigma_{W,jk}^2$ are drawn from $U(V_L, V_H)$. The empirical distributions of the estimated and simulated variances appear to have comparable features.

For a set of generated proficiency scores the jackknife method, with the replicate weights from the survey, and the model-based methods were applied. The following estimators were evaluated:

- the weighted mean;

Table 7: Summary of simulation of model-based estimators.

Estimator	Minimum	Mean	Median	Maximum	St. dev.	Deg. fr.
Wtd mean	248.144	250.273	250.292	252.672	1.197	
Jack. mean	248.144	250.275	250.285	252.655	1.200	
Jack. var.	0.720	1.351	1.328	2.251	0.340	31.6
CI var	0.657	1.406	1.372	2.265	0.320	38.7
CI σ^2_B	32.231	100.285	97.044	180.142	29.092	
CII var	0.657	1.402	1.365	2.257	0.319	38.6
CII σ^2_B	32.231	99.866	96.157	179.413	29.003	
CIII var	0.633	1.412	1.396	2.367	0.324	38.0
CIII σ^2_B	30.043	100.815	99.150	189.343	29.445	
CR var	0.631	1.403	1.378	2.257	0.319	38.7
CR σ^2_B	29.829	100.031	97.550	179.392	29.005	
C100 var	0.639	1.418	1.392	2.288	0.317	40.1
C100 σ^2_B	30.550	101.333	98.725	182.211	28.808	
REML var	0.631	1.403	1.378	2.257	0.319	38.7
REML σ^2_B	29.830	100.032	97.550	179.393	29.005	
σ^2_W mean	609.969	667.077	667.641	712.931	21.561	

Notes: The estimators of the sampling variance of the weighted mean are denoted by the method (for instance, CR) and the symbol 'var' in the first column. The estimators of the between-cluster variance are denoted by the method and the symbol σ^2_B. The estimator of the common within-cluster variance is given in the last row of the table (for all methods separate within-cluster variances are estimated for each cluster). The group means, common to the set of simulations, were generated from $\mathcal{N}(250, 10)$. The within-cluster deviations were generated from centered normal distributions with standard deviations drawn from $\mathcal{U}(10, 40)$, the between-cluster deviations were generated from $\mathcal{N}(0, 10)$. All the random draws were mutually independent. The standard deviations were common to the simulations, but the (random) deviations were drawn independently for each replicate. One hundred replicates were simulated.

- the jackknife mean;
- the within-cluster variances;
- the between-cluster estimates; $\sigma^2_{C,I}$, $\sigma^2_{C,II}$, $\sigma^2_{C,III}$, $\sigma^2_{C,R}$, $\sigma^2_{C,100}$, calculated as $\sigma^2_{C,R}$ with weights based on $\sigma^2_B = 100$, and $\sigma^2_{C,ML}$, an iteration of the weighted Fisher scoring algorithm described in Section 3.9;
- the estimated sampling variances of the estimators of the between-cluster variance.

Table 7 contains a summary of a set of simulations for the population mean. For each estimator (a row of the table), the minimum, mean, median, and maximum realized value are given, as well as the standard deviation of the realized values. The quantities in the extreme right column (degrees of freedom)

are discussed below.

The weighted mean (ratio) and the jackknife estimators of the population mean are almost identical. Their apparent bias (from the superpopulation mean of 250) is due to the uncertainty associated with the simulated group-level deviations which are constant across the simulations. The true variance of the weighted mean estimator can be calculated by substituting the generated values of the variances $\sigma^2_{W\,jk}$ in the equation (18); its value, 1.389, implies that the jackknife estimator of the sampling variance (mean 1.35, median 1.33) has a negative bias, while the model-based estimators have positive biases. However, the contributions of these biases to the mean squared errors of the estimated variance are negligible. Also, the bias in estimation of the between-cluster variance ($\sigma^2_B = 100$) is trivial in comparison with the sampling variance of the estimator.

The model-based methods CI, CH, CHI, and CR, were introduced above. The method denoted as C100 in Table 7 is analogous to CR, but the weights u_{jk} are calculated using $\sigma^2_B = 100$. This means that the weights u_{jk} would be optimal if $\sigma^2_{W\,jk}$ were equal to their estimates, and $\sigma^2_B = 100$. Further, REML stands for the weighted version of the restricted maximum likelihood method described in Section 3.9. Since a good starting solution is used (σ^2_B from method CR) only one iteration of the Fisher scoring algorithm is applied; preliminary exploration showed that further iterations change the value of the estimate of σ^2_B by less than 0.1.

All the estimators of the sampling variance of the weighted mean appear to be downward biased; the observed variance of the estimator of the mean is $1.197^2 = 1.428$. The estimator C100 nearly matches this value (its observed mean is 1.418), and the other model-based estimators are only marginally more biased. The jackknife estimator of the sampling variance has by far the largest bias. Furthermore, the observed standard deviation of the jackknife estimator is slightly higher than its model-based counterparts, so that its mean squared error is also the highest. The comparisons of the stardard errors (square roots of the sampling variances) lead to the same conclusions.

Thus, there is little to choose between the model-based estimators of the sampling variance, perhaps with the exception of the estimator C100 which assumes known between-cluster variance. It is encouraging, though, that not knowing this variance causes only marginal loss of efficiency. The model-based estimators have almost identical distributions and they are also very highly correlated. The additional computation involved in the method CR makes only a marginal contribution to the efficiency.

However, for several simulations the differences between the estimates are considerable, as can be seen in the pairwise plots of the sets of estimates in Figure 7. The methods CR and C100 are not represented in the plots so as to achieve higher resolution and clarity.

The (model-based) estimators of σ^2_B are also very similar and mutually highly correlated. The estimators appear to be unbiased, although their distributions are somewhat skewed. Note the substantial uncertainty in their estimation; their observed standard deviations are around 29. It is surprising that the C100 estimator has the largest bias; however, its observed standard deviation (28.81) and its mean squared error (28.84) are the smallest.

An intuitively appealing way of comparing the efficiency of the estimators of the sampling variance is

BEST COPY AVAILABLE

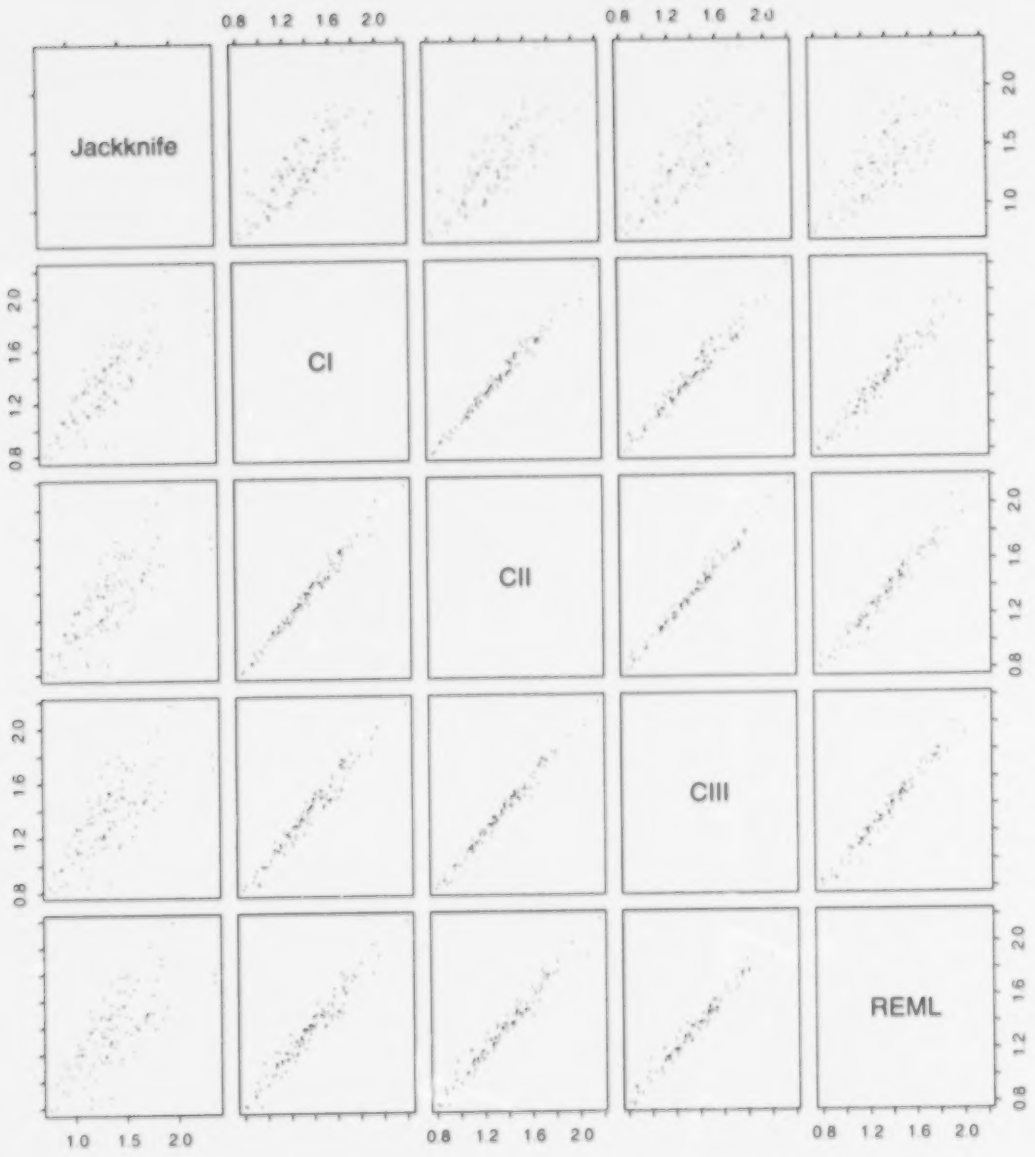

Figure 7: Pairwise plots of the simulated estimates of sampling variance using jackknife and methods CI, CII, CIII, and REML.

by fitting χ^2 distributions to the empirical distributions of the estimators. Then more degrees of freedom of the fitted χ^2 distribution imply higher efficiency. The method of moments estimator of the degrees of freedom is given by the equation

$$\widehat{\text{d.f.}} = \frac{2 \times (\text{mean estimate of variance})^2}{\text{variance of the m.s.e.}} \tag{33}$$

These degrees of freedom are given in the extreme right column of Table 7. In evaluation of this estimator we ignore the bias of the variance estimator. Not knowing σ_B^2 is associated with loss of up to 2 degrees of freedom, and using the jackknife method with an additional loss of 6.5 degrees of freedom.

The model-based estimators are mutually highly correlated; the correlations among the estimators CH, CHH, CR, C100, and REML are 0.990 or higher, the correlations of these estimators with the estimator CI is 0.981 or higher. The correlations of the jackknife estimator with the model-based estimators are 0.800 or higher; the highest correlation is with estimator CI (0.865).

The set of 100 simulations was repeated with the same simulation parameters using a different sample of simulated group deviations. Although substantially different values of the observed means and standard deviations were obtained, their pattern, and conclusions about efficiency, were the same. For instance, in one case the jackknife estimator of the sampling variance of the weighted mean had the least bias, but in all cases it was associated with 20 – 39 per cent loss of efficiency vis-à-vis either of the model-based estimators.

In conclusion, the jackknife method does not deliver on its promise of unbiased estimation of the sampling variance, and the efficiency of this estimator is also inferior to its model-based counterparts. The model-based methods do not involve any appreciable loss of efficiency due to unknown between-cluster variance. There appears to be little payoff for the set of 56 replicate weights required for the jackknife estimation, which are dispensed with in the model-based methods.

Generalizability of these conclusions was explored by further simulations using different parameter values and different datasets. The jackknife estimator of the sampling variance is particularly vulnerable when the between-cluster variance σ_B^2 is large; the loss of efficiency for $\sigma_B^2 = 150$ is about 50 percent, while for $\sigma_B^2 = 50$ it is around 20 per cent. The choice of the distribution for the between-group differences does not appear to affect the properties of the estimators. The estimators of σ_B^2 have negligible bias, and, agreeing with intuition, have sampling variances increasing with σ_B^2.

4.1 Estimators for subpopulations

Of particular importance are the relative performances of the studied estimators for smaller samples or subpopulations. We describe in details the simulations for the subpopulation of Hispanic students (response 3 to item 28) in the New Jersey sample. There are 363 Hispanic students in the sample; they are located in 84 clusters within 52 groups in the sample. The distribution of the Hispanic students across the clusters is compactly summarized in Table 8. A large number of clusters contain only one to three Hispanic students, while in a few clusters Hispanic students form a majority.

4.1

Table 8: Distribution of Hispanic students in the New Jersey sample.

	Students within clusters															
Students	1	2	3	4	5	6	7	8	10	11	12	13	14	16	17	22
Clusters	27	16	10	3	7	5	3	1	1	2	2	2	1	1	2	1

Note: The second row contains the numbers of clusters which have the numbers of students given in the same column. For instance, 27 clusters have one Hispanic student each.

Several values of the model-based estimators of σ_B^2 are negative (up to 12 per 100 simulated values), and, for the purpose of calculating the weights u_{jk}, they are truncated to zero throughout. The results of the simulations are given in Table 9 in the same format as Table 7.

The observed variance of the weighted means is equal to 5.51. The estimators CR and REML, come closest to matching this value, followed by C100 which assumes known variance σ_B^2 equal to the corresponding simulation parameter. The price for unbiasedness is very high, though; both CR and REML have much higher mean squared errors than the jackknife (2.20), or the other model-based estimators (1.88 for C100, 1.95 for CH, 2.06 for CI, and 2.12 for CHH). The degrees of freedom loosely reflect the efficiency of the estimators, although upward biased estimators appear in a somewhat better light. The difference of 0.7 degrees of freedom between the estimators C100 and CH, that can be attributed to information about σ_B^2, appears to be trivial in comparison with the differences among the least biased and jackknife estimators on one hand, and the estimators CI, CH, and CHH on the other hand.

The loss of efficiency due to not knowing σ_B^2 is quite moderate, even though estimation of σ_B^2 is associated with a lot of uncertainty. However, if an incorrect value of σ_B^2 is assumed for the estimator C100, a substantial loss of efficiency is incurred. For example, assuming that $\sigma_B^2 = 150$ when in fact $\sigma_B^2 = 100$ yields a severely biased estimator with mean squared error comparable to the jackknife. In small subsamples the sampling variance of the weighted mean is more strongly influenced by σ_B^2 than in the entire sample. For example, the sampling variance of the weighted mean simulated using $\sigma_B^2 = 70$ is about one half of the sampling variance simulated using $\sigma_B^2 = 150$.

In all datasets the assumption of equal within-cluster variances (σ_W^2) is associated with substantial loss of efficiency.

These observed properties were confirmed in simulations based on several other subsamples with sizes 130 – 500. In general, the jackknife estimator is biased and its sampling variance is larger than that of the model-based estimators, with occasional exception of CR. The estimator CHH is the most efficient one for some subsamples (following C100), but performs rather poorly for others, though never worse than the jackknife. The performance of the estimators CI and CH is much more consistent; CH is uniformly more efficient than CI, but the difference is unimportant in comparison with the improvement these estimators represent over the other methods. The jackknife is least competitive for the smallest datasets (losses in efficiency of up to 45 per cent) and, ironically, for the entire sample. Inefficiency of the jackknife for larger

Table 9: Summary of simulation of model-based estimators for Hispanic students.

Estimator	Minimum	Mean	Median	Maximum	St. dev.	Deg. fr.
Wtd mean	243.00	248.75	248.78	255.99	2.35	
Jack. mean	242.94	248.76	248.78	255.98	2.34	
Jack. var.	1.86	4.92	4.58	13.26	2.11	10.9
CI var	2.19	4.89	4.62	11.30	1.96	12.5
CI σ^2_B	0	112.59	97.34	381.43	80.48	
CII var	1.88	5.33	4.90	10.57	1.94	15.1
CII σ^2_B	0	89.47	68.11	383.14	77.95	
CIII var	2.40	5.73	5.34	11.20	2.10	15.0
CIII σ^2_B	0	127.78	111.45	383.04	85.44	
CR var	1.93	5.42	4.78	16.64	2.84	7.3
CR σ^2_B	0	114.91	96.78	407.29	97.20	
C100 var	3.73	5.23	4.90	12.69	1.86	15.8
C100 σ^2_B	71.18	109.90	99.20	430.66	84.61	
REML var	1.93	5.42	4.78	16.63	2.83	7.3
REML σ^2_B	0	114.90	96.87	438.36	97.24	
σ^2_W mean	413.20	688.69	669.80	1049.21	106.12	

Notes: The same notation and layout is used as in Table 7. Three hundred replicates were simulated, using the simulation parameters given in Table 7.

samples may be due to not using within-cluster information.

5 Smoothing techniques

The model-based method of estimation of the standard error of the mean proficiency for a subpopulation involves estimation of the cluster- and student-level variance components σ^2_B and $\{\sigma^2_{W,jk}\}$. For small subsamples, especially those with only a few strata represented by more than one large cluster, the estimates of these variances have large sampling variances. Clearly, estimation of these variances is the Achilles heel of the model-based methods; it is exceedingly inefficient when a large number of subsamples is analyzed because information about the variances contained in the analyzed subsample could be complemented by the other subsamples.

Although it is not reasonable to assume that all the subsamples have the same between-cluster variance σ^2_B, suitably selected sets of subsamples may share a common variance σ^2_B. Then estimation of σ^2_B can be strengthened by averaging the estimates of σ^2_B across the subsamples. In this process of averaging more weight can be given to larger subsamples. Also, the weights may vary depending on the analyzed

subsample. Schemes motivated by shrinkage estimation or empirical Bayes methods may be particularly useful. In a typical such scheme each variance for a subsample is estimated as the weighted mean of the variances for the entire set of subsamples, with the weights associated with each subsample held constant, except for the analyzed subsample which is given more weight.

Care has to be exercised in averaging the estimated within-cluster variances for a cluster because subsamples may have substantially different within-cluster variances. Each cluster within a subsample can be considered as an informatively selected subsample. The change in the within-cluster variances is greater the more closely the variable on which the selection is based is associated with proficiency. In any case, estimation of the within-cluster variance is simple (based on the sum of squares $v_{A,jk}$), and a variety of schemes of pooling information across subsamples and/or clusters can be devised.

The influence of the sampling errors associated with estimation of the variance components can be further reduced by the following scheme. The sampling variance of the estimator of a subpopulation mean is a linear function of the variance components:

$$\mathrm{var}(\mu_a) \;=\; S_{a,2}\sigma^2_B + \sum_{jk} S_{a,1jk}\sigma^2_{W,jk}\,, \tag{34}$$

where $S_{a,2}$ and $S_{a,1jk}$ are functions of the sampling weights;

$$S_{a,2} \;=\; \frac{1}{n_{a,B}} \;=\; \frac{W^2_a}{\sum_{jk} W^2_{a,jk}}$$

$$S_{a,1jk} \;=\; \frac{\sum_i w^2_{ijk}/n_{a,B}}{\sum_{jk} W^2_{a,jk}} \;=\; \frac{W^2_a \sum_i w^2_{ijk}}{(\sum_{jk} W^2_{a,jk})^2}\,.$$

The subscript $a = 1, \ldots, A$ is added to several quantities in (34) and throughout this section to emphasize their dependence on the analyzed subsample (dataset). For the estimated sampling variances $\widehat{\mathrm{var}}(\mu_a)$ for subpopulations a we consider the regression equation

$$\widehat{\mathrm{var}}(\mu_a) \;=\; S_{a,2}\sigma^2_B + \sum_{jk} S_{a,1jk}\sigma^2_{W,jk} + \varepsilon_a\,, \tag{35}$$

where ε_a, $a = 1, \ldots, A$, are subpopulation-specific random terms and σ^2_B and $\{\sigma^2_{W,jk}\}$ are sets of variances common across the subpopulations. The term ε_a consists of two components: the error of estimation, $\widehat{\mathrm{var}}(\mu_a) - \mathrm{var}(\mu)$, and the model deviation $\mathrm{var}(\mu) - S_2\sigma^2_{a,B} - \sum_{jk} S_{a,1jk}\sigma^2_{W,jk}$. The 'regression' parameters σ^2_B and $\sigma^2_{W,jk}$ may be assumed known or unknown. In the latter case they can be estimated by standard regression methods.

Thus, given a set of estimates of the sampling variances $\mathrm{var}(\mu_a)$, and assuming common variance components across the subsamples $\{a\}$, these estimates can be 'improved', or smoothed, by fitting the linear model (35), and declaring the fitted values

$$\widehat{\widehat{\mathrm{var}}}(\mu_a) \;=\; S_{a,2}\sigma^2_B + \sum_{jk} S_{a,1jk}\sigma^2_{W,jk}\,, \tag{36}$$

where the estimates of the variance components are obtained by a weighted least squares fit with weights reflecting differential precision of the estimators $\widehat{\text{var}}(\mu_a)$. Since there are a large number of clusters it is expedient to substitute the estimates $\hat{\sigma}^2_{W\,jk}$ for the corresponding variances in (34) thus obtaining a simple regression on σ^2_B with no intercept. To take account of differential precision of the estimated variances $\widehat{\text{var}}(\mu)$ (estimated, say, by jackknife) suitable regression weights, such as the sample size, cluster sample size, or their linear combination can be applied. Stability of the estimate of the common between-cluster variance σ^2_B can be explored by varying (perturbing) the regression weights. An important diagnostic check for appropriateness of the model in (34) is that the regression intercept, if estimated, is close to zero. The smoothed estimate of the sampling variance is the fitted value (36), where the variances are either estimated exclusively from the subsample to which they refer, by a regression (common to all subsamples), or as a compromise of these two approaches (such as an empirical Bayes approach).

Extensions and several adaptations of this approach are easy to devise, for example, by introducing different variances σ^2_B for disjoint subsets of subsamples. The regression equation in (34) can be suplemented by other data summaries (not only functions of the sampling weights), as well as by an intercept term, thus obtaining a better fit, although the original interpretation in terms of a common between-cluster variance would no longer apply.

An important concern pertains to normality and homogeneity of the 'error' terms ε_a. Instead of the regression of $z_a = \widehat{\text{var}}(\mu) - \sum_{rj} S_{a\,1jk}\hat{\sigma}^2_{W\,jk}$ on $S_{a\,2}$, we may consider the regression

$$z_a/S_{a\,2} \;=\; \sigma^2_B + \varepsilon^*_a\,, \tag{37}$$

in which the assumption of i.i.d. for $\varepsilon^*_a = \varepsilon_a/S_{a\,2}$ may be more palatable. Now the common variance σ^2_B is estimated as the mean of the quantities $z_a/S_{a\,2}$. As an alternative, a suitable transformation of z_a can be applied; in particular, if $\min(z_a)$ is positive, log-transformation leads to the *geometric* average of z_a as an estimator of σ^2_B. A transformation may also be applied to $z_a/S_{a\,2}$.

Of course, the method outlined in this section can be applied to another model for the outcome, which would lead to a relationship between sampling variance and a different set of summaries of the dataset. The method requires an estimator of the sampling variances. The jackknife or a model-based estimator can be used, and it is then improved by the smoothing.

An important issue in application of these methods is identification of subsets with equal (homogeneous) within- and between-cluster variances. Detailed understanding of the educational, social, behavioural, and economic processes relevant to the target population may promote an intelligent choice in this myriad of smoothing schemes. An interesting option is that of combining the estimates of the variances based on a subsample with the estimates of variances (their means) from the other subsamples. The mixing proportions should depend on the (effective) sample size(s) of the analyzed subsample. Such a scheme, motivated by empirical Bayes methods, has a great potential but requires careful experimentation and fine-tuning which are beyond the scope of this project.

6 Regression with survey data

Ordinary regression provides an easy to interpret assessment of association of one variable (the *response*) with a set of other (*explanatory*) variables. The standard least squares method for estimation of this regression function is applicable when the following assumptions are satisfied: the values of the explanatory variables are under control of the experimenter/analyst, the outcomes are generated by a process which assigns values of the response conditionally independent and normally distributed with constant (conditional) variance given the relationship to the explanatory variables (regression function), and the regression function is linear in the explanatory variables. In general, the assumptions of conditional independence and of non-informative selection of subjects are crucial for validity of the inference based on the regression estimator.

In regression analysis using survey data there are two distinct challenges: definition of the estimated quantity, and taking account of the features of the sampling design. For conceptual clarity we consider first the hypothetical situation in which the values of the response Y and of the (single) explanatory variable X, denoted respectively by Y_i and X_i, $i = 1, \ldots, N$, are available for the entire population. With such data we would calculate the regression slope as

$$\beta = \frac{\sum_i (X_i - \bar{X})Y_i}{\sum_i (X_i - \bar{X})^2} = \frac{N^{-1}\sum_i X_i Y_i - \bar{X}\bar{Y}}{N^{-1}\sum_i X_i^2 - \bar{X}^2}, \tag{38}$$

where $\bar{X} = N^{-1}\sum_i X_i$ and $\bar{Y} = N^{-1}\sum_i Y_i$ are the respective population means of X and Y; all the summations are over $i = 1, \ldots, N$. Since our inference is conditional on the target population, we regard β in (38) as an unknown constant.

As an alternative, we may consider a construct (latent) variable Y^*, observed or measured indirectly and subject to random deviation (error) by the variable Y, and suppose that Y^* is linearly related to X:

$$Y_i^* = \alpha^* + \beta^* X_i .$$

If the values of X_i and Y_i were available for the entire population β would, under certain standard assumptions, be a 'good' estimator of β^*.

The (residual) variance of the deviations $Y_i - Y_i^*$ is estimated as

$$\sigma^2 = \sigma_*^2 = \frac{1}{N-2}\sum_i (Y_i - \alpha - \beta X_i)^2, \tag{39}$$

where $\alpha = \bar{Y} - \bar{X}\beta$. We emphasize that the variance σ^2 in (39) is a constant or an estimator, depending on the adopted perspective. The variance of β, as an estimator of β^*, is $\sigma_*^2 / \{\sum_i (X_i - \bar{X})^2\}$, and would itself be estimated by

$$\sigma^2 \left\{ \sum_i (X_i - \bar{X})^2 \right\}^{-1} . \tag{40}$$

We consider estimation of the quantities (38) – (40) based on a stratified clustered sample from the target population. Our approach is based on estimation of the population means $N^{-1}\sum_i X_i Y_i$, $N^{-1}\sum_i X_i^2$, $N^{-1}\sum_i Y_i^2$, $\bar{X}$, and $\bar{Y}$, for which methods discussed in Section 3.4 are applicable.

As the estimator of the population summary $N^{-1}\sum_i X_i Y_i$ we use the statistic

$$\overline{xy} \;=\; \frac{\sum_{ijk} w_{ijk} x_{ijk} y_{ijk}}{\sum_{ijk} w_{ijk}}, \tag{41}$$

where the summations are over all the *sampled* elementary units (students i in cluster j in group k). In order to find the (approximate) distribution of the estimators of (38) – (40) based on a random sample, we require the covariance structure of the estimators of the population means. The statistic $\overline{xy}$ is the estimator (12) applied to the product XY. Its sampling mean and variance are derived in Section 3.4.

For the product XY we consider the estimator

$$xy \;=\; \frac{\sum_{ijk} w_{ijk} x_{ijk} \sum_{ijk} w_{ijk} y_{ijk}}{\left(\sum_{ijk} w_{ijk}\right)^2};$$

its expectation is

$$\mathbf{E}(xy) \;=\; \mathrm{cov}(x, y) + \mathbf{E}(x)\mathbf{E}(y).$$

The covariance can be expressed in terms of variances:

$$\mathrm{cov}(x, y) \;=\; \frac{1}{2}\left\{\mathrm{var}(X + Y) - \mathrm{var}(X) - \mathrm{var}(Y)\right\}. \tag{42}$$

Note that xy is an unbiased estimator of XY only when x and y are uncorrelated. This is unlikely, though, when X and Y are associated. The bias is small when $|\,\mathrm{cov}(xy)\,|$ is much smaller than XY.

The sampling variance of xy is

$$\mathrm{var}(xy) \;=\; \mathrm{cov}(x^2, y^2) + \mathbf{E}(x^2)\mathbf{E}(y^2) - \{\mathbf{E}(xy)\}^2. \tag{43}$$

Evaluation of the covariance $\mathrm{cov}(x^2, y^2)$ is in general not straightforward. If the assumptions of normality of the means x and y are adopted, then this covariance is a function of means, the covariance, and the variances of x and y. The derivation, given below, uses the properties of the conditional distributions under normality assumptions.

The conditional distribution of x given y is

$$\mathcal{N}\left\{\mu_x + \frac{\sigma_{xy}}{\sigma_y^2}(y - \mu_y),\; \sigma_x^2 - \frac{\sigma_{xy}^2}{\sigma_y^2}\right\},$$

where μ and σ, with the appropriate subscript(-s), stand for the expectations and (co-)variances of x and y. Now

$$\mathbf{E}(x^2 \mid y) \;=\; \sigma_x^2 - \frac{\sigma_{xy}^2}{\sigma_y^2} + \mu_x^2 + 2\mu_x \frac{\sigma_{xy}}{\sigma_y^2}(y - \mu_y) + \frac{\sigma_{xy}^2}{\sigma_y^4}(y - \mu_y)^2,$$

and hence, taking expectation over y and assuming that it is normally distributed, we obtain

$$\begin{aligned}
\mathbf{E}\,(x^2 y^2) \;&=\; \mathbf{E}_y\left\{y^2 \mathbf{E}(x^2 \mid y)\right\} \\
&=\; (\sigma_y^2 + \mu_y^2)\left(\sigma_x^2 - \frac{\sigma_{xy}^2}{\sigma_y^2} + \mu_x^2\right) + 2\mu_x\mu_y\frac{\sigma_{xy}}{\sigma_y^2} + 3\sigma_{xy}^2\sigma_y^2 + \mu_y^2\frac{\sigma_{xy}^2}{\sigma_y^2} \\
&=\; (\sigma_x^2 + \mu_x^2)(\sigma_y^2 + \mu_y^2) + 2\sigma_{xy}(\sigma_{xy} + \mu_x\mu_y).
\end{aligned} \tag{44}$$

The other terms in the estimator

$$\beta = \frac{\sum_{ijk} w_{ijk} x_{ijk} y_{ijk} - W\bar{x}\bar{y}}{\sum_{ijk} w_{ijk} x_{ijk}^2 - W\bar{x}^2} \tag{45}$$

(W is the sample total of sampling weights) can be obtained from the sampling means and variances of x, y, and their covariance.

The expectations and variances of the numerator and denominator in (45) are found by the following method. Let x_a, x_b, and x_c be three normally distributed random variables with respective means μ_a, μ_b, and μ_c, and (co-)variances Σ (subscripts a, b, and c). Then

$$\begin{aligned}
\mathrm{var}(x_a - x_b x_c) &= \mathbf{E}\{(x_a - x_b x_c)^2\} - \{\mathbf{E}(x_a - x_b x_c)\}^2 \\
&\quad \Sigma_{aa} + \mu_a^2 + \Sigma_{bb}\Sigma_{cc} + 2\Sigma_{bc}^2 + \mu_b^2\Sigma_{cc} + \mu_c^2\Sigma_{bb} + \mu_b^2\mu_c^2 \\
&\quad - 2\mu_a\Sigma_{bc} - 2\mu_b\Sigma_{ac} - 2\mu_c\Sigma_{ab} - 2\mu_a\mu_b\mu_c + 4\mu_b\mu_c\Sigma_{bc} \\
&\quad - (\mu_a^2 + \Sigma_{bc}^2 + \mu_b^2\mu_c^2 - 2\mu_a\mu_b\mu_c - 2\mu_a\Sigma_{bc} + 2\mu_b\mu_c\Sigma_{bc}) \\
&= \Sigma_{aa} + \Sigma_{bb}\Sigma_{cc} + \Sigma_{bc}^2 + \mu_b^2\Sigma_{cc} + \mu_c^2\Sigma_{bb} \\
&\quad -2\mu_b\Sigma_{ac} - 2\mu_c\Sigma_{ab} + 2\mu_b\mu_c\Sigma_{bc}\,.
\end{aligned} \tag{46}$$

The equation for the denominator is obtained from (46) by setting $x_b = x_c$ ($\Sigma_{bb} = \Sigma_{bc} = \Sigma_{cc}$, $\Sigma_{ab} = \Sigma_{ac}$, and $\mu_b = \mu_c$):

$$\mathrm{var}(x_a - x_b^2) = \Sigma_{aa} + 2\Sigma_{bb}^2 + 4\mu_b^2\Sigma_{bb} - 4\mu_b\Sigma_{ab}\,.$$

Finally, to determine the moments of the ratio in (45) we require the covariance of the numerator and denominator. An exact method appears to require considerable effort since moments of various non-linear functions of the data are required. On the other hand, especially if β is positive, the correlation of the numerator and denominator is likely to be very high, and the distribution of the ratio can be estimated by imputing one or several realistic values of this correlation.

We write the estimator (45) as

$$\beta = \frac{u_1 + \gamma_1}{u_2 + \gamma_2} \tag{47}$$

where u_1 and u_2 are constants, the respective expectations of the numerator and denominator in (45), and γ_1 and γ_2 are centered random variables with respective variances σ_1^2 and σ_2^2 and covariance σ_{12}. Let $\rho_{12} = \sigma_{12}/\sqrt{\sigma_1^2\sigma_2^2}$ be the correlation of γ_1 and γ_2. Suppose $|\gamma_h|$ is much smaller than u_h with high probability ($h = 1, 2$), that is, σ_h is smaller than u_h for $h = 1, 2$. Then from the expansion

$$\begin{aligned}
\beta &= \left(\frac{u_1}{u_2} + \frac{\gamma_1}{u_2}\right)\left(1 - \frac{\gamma_2}{u_2} + \frac{\gamma_2^2}{u_2^2} - \ldots\right) \\
&\approx \frac{u_1}{u_2} + \frac{\gamma_1}{u_2} - \frac{\gamma_2 u_1}{u_2^2} - \frac{\gamma_1\gamma_2}{u_2^2} + \frac{\gamma_2^2 u_1}{u_2^3}
\end{aligned}$$

(ignoring the higher-order terms), we have

$$\begin{aligned} \mathbf{E}(\beta) &\approx \frac{u_1}{u_2} - \frac{\sigma_{12}}{u_2^2} + \frac{\sigma_2^2 u_1}{u_2^3} \\ \mathrm{var}(\beta) &\approx \frac{\sigma_1^2}{u_2^2} + \frac{\sigma_2^2 u_1^2}{u_2^4} - 2\sigma_{12}\frac{u_1}{u_2^3} \end{aligned} \tag{48}$$

These equations can be supplemented by further terms involving higher moments of the random variables γ_1 and γ_2. The values of these moments are not determined by the variance matrix of (γ_1, γ_2) unless normality is assumed. However, the assumptions of normality are palatable for large sample sizes.

Note that unbiasedness of the numerator and denominator in (45), as estimators of their respective population counterparts, does not imply unbiasedness of β, not even when $\sigma_{12} = 0$.

6.1 Residual variance

The residual variance σ^2 can be estimated by the same approach. Ignoring the degrees of freedom in regression for the population, $N/(N-2) \doteq 1$, we have

$$\sigma^2 = \overline{Y^2} - \bar{Y}^2 - \frac{\left(\overline{XY} - \bar{X}\bar{Y}\right)^2}{\overline{X^2} - \bar{X}^2}, \tag{49}$$

where the bar over variables X and Y and their functions stands for corresponding population mean. The naive estimator of σ^2 is constructed by replacing each population mean in (49) by its (jackknife or model-based) estimator. The mean square $\overline{Y^2} - \bar{Y}^2$ is estimated without bias by $\overline{y^2} - \bar{y}^2 - \widehat{\mathrm{var}}(\bar{y})$. The denominator of the fraction in (49) is estimated by $\overline{x^2} - \bar{x}^2 - \widehat{\mathrm{var}}(\bar{x})$, and the numerator by

$$(\overline{xy} - \bar{x}\bar{y})^2 - \widehat{\mathrm{var}}(\overline{xy} - \bar{x}\bar{y}).$$

To obtain the expectation of this estimator we consider the expectations of the numerator and denominator in (49); the former is

$$\mathbf{E}\left\{(\overline{xy} - \bar{x}\bar{y})^2\right\} = \mathrm{var}(\overline{xy} - \bar{x}\bar{y}) + \left\{\mathbf{E}(\overline{xy} - \bar{x}\bar{y})\right\}, \tag{50}$$

and the latter is derived in complete analogy with the mean square for Y.

6.2 Implementation

The estimation procedure described above requires estimation of the sampling variance matrix of the means of X, Y, X^2, XY, and Y^2. This can be accomplished by jackknife, the model-based, or, in principle, any other method. The extension of both for multivariate statistics is discussed in Section 3.8. The estimated moments are then substituted for the 'true' moments in (46) applied for the numerator and denominator of the regression estimator β, the estimators of (39) and (40). The estimator β can be (approximately) corrected for bias using equation (48).

Note that the method described in this section is applicable for any sampling design since it is based on the expectations and covariance structure of the population means of certain variables.

6.3 Regression with jackknife

The jackknife method discussed in Section 3.3 has a straightforward extension to ordinary regression. In essence, we replace the mean as the 'parent' method by the weighted regression, and subject each component of the regression vector estimate (the intercept and slope in the case of simple regression), as well as the residual variance, to the jackknife estimation.

Thus, in NAEP Trial State Assessment an ordinary regression is fitted for combination of each set of replicate weights (and the final sampling weights) and plausible values, a total of $57 \times 5 = 285$ regressions. The sets of estimates of the regression parameters and of the residual variances are then summarized to obtain the jackknife estimates of these parameters. Although relatively easy to implement this procedure demands a lot of computing time without necessarily providing an efficient estimator of the standard errors. A short-cut, using the jackknife as described above for the first plausible value, and estimating the residual variance from the regressions with the final weights for the other four plausible values, is used in operation. This way only 61 regressions have to be fitted.

6.4 Example

For illustration we construct a regressor variable as the total of all the non-missing responses to items 231 – 238 (*'In Math class how often do you ...'*). Each of these items is scored on the Likert scale (1 – 5, ordinal). Data from 2171 students (80.1 per cent of the sample) from 104 clusters, who responded to each item are used in the analysis. The jackknife analysis, involving $57 \times 5 = 285$ weighted least squares regression fits, is summarized in the top panel of Table 10. The jackknife estimates of the intercept, slope, and residual variance are given for each plausible value, and the estimates for the proficiency scores are given in the right-most column.

The results of the model-based regression method are given in the bottom panel of Table 10. They are in close agreement with the jackknife method. The estimated correlation of the numerator and denominator of β in (48) is equal to 0.40, but even for imputed correlations of zero and unity the results are not substantially different. Table 11 displays the results for the proficiency scores, summarizing the analyses for each plausible value. To explore the influence of the imputed correlation, the results are given for correlations $0, 0.2, \ldots, 1$. The standard errors of the slope estimator are affected a great deal by the choice of the correlation, but for the estimates (intercept, slope, and residual variance) the choice of the correlation is not critical. This suggests a simplification of the Taylor expansion method presented in Section 6.

6.5 Multivariate and multilevel regression

The model-based method for simple regression can serve as an outline for extensions to multivariate and multilevel regression. In the former, the population regression parameter is defined as

Table 10: Regression analysis using the jackknife and model-based methods; New Jersey data.

Parameter	Plausible value 1	2	3	4	5	Prof. value
Jackknife						
Intercept	230.03	230.02	229.63	230.33	226.66	229.34
Slope	1.403 (0.227)	1.401 (0.218)	1.414 (0.218)	1.390 (0.225)	1.524 (0.235)	1.426 (0.235)
Res. variance	1042.05	1031.51	1015.88	1043.62	1048.37	1036.47
Model-based method						
Intercept	230.03	229.02	229.62	230.16	226.61	229.28
Slope	1.407 (0.224)	1.406 (0.216)	1.418 (0.219)	1.395 (0.223)	1.528 (0.236)	1.432 (0.232)
Res. variance	1040.12	1028.51	1013.17	1040.37	1048.33	1036.44

Note: The estimated standard errors are given in parentheses.

$$\boldsymbol{\beta} = \left(\mathbf{X}^\top \mathbf{X}\right)^{-1} \mathbf{X}^\top \mathbf{Y}, \tag{51}$$

where $\mathbf{X}$ is the $N \times p$ (population) matrix of the regressors (the population design matrix), and $\mathbf{Y}$ is the $N \times 1$ vector of the outcomes for the population. Each total of crossproducts in (51) can be estimated by the corresponding ratio estimator and the sample covariances can be obtained from the sampling variances using the formula

$$\widehat{\mathrm{cov}}(\overline{x_1 x_2}, \overline{x_3 x_4}) = \frac{\widehat{\mathrm{var}}(\overline{x_1 x_2} + \overline{x_3 x_4}) - \widehat{\mathrm{var}}(\overline{x_1 x_2}) - \widehat{\mathrm{var}}(\overline{x_3 x_4})}{2},$$

where $\overline{x_1 x_2}$ denotes the ratio estimator of the population mean product of $X_1 X_2$. The expectation and the variance matrix of the sample counterpart of (51),

$$\beta = \left(\overline{\mathbf{x}^\top \mathbf{x}}\right)^{-1} \overline{\mathbf{x}^\top \mathbf{y}},$$

can be approximated by the multivariate version of the delta method. Let

$$\hat{\boldsymbol{\beta}} = (\mathbf{u}_2 + \boldsymbol{\varepsilon}_2)^{-1}(\mathbf{u}_1 + \boldsymbol{\varepsilon}_1),$$

where $\mathbf{u}_h$ are the expectations and $\boldsymbol{\varepsilon}_h$ the deviations from the expectations of the numerator ($h = 1$) and denominator ($h = 2$) in (51). We denote the covariance matrices for row h of $\boldsymbol{\varepsilon}_2$ with $\boldsymbol{\varepsilon}_1$ by $\boldsymbol{\Sigma}_{21,h} = \mathrm{cov}(\boldsymbol{\varepsilon}_{2,h}, \boldsymbol{\varepsilon}_1)$, for two rows of $\boldsymbol{\varepsilon}_2$ by $\boldsymbol{\Sigma}_{2,hh'} = \mathrm{cov}(\boldsymbol{\varepsilon}_{2,h}, \boldsymbol{\varepsilon}_{2,h'})$, and set $\boldsymbol{\Sigma}_1 = \mathrm{var}(\boldsymbol{\varepsilon}_1)$.

Assuming that $|\boldsymbol{\varepsilon}_2|$ is much smaller than $\mathbf{u}_2$ we have the expansion

Table 11: Regression analysis of the proficiency on a constructed variable using the model-based method; New Jersey data.

	Correlation					
Parameter	0	0.2	0.4	0.6	0.8	1.0
Intercept	229.45	229.37	229.29	229.21	229.13	229.04
Slope	1.426 (0.252)	1.430 (0.242)	1.432 (0.232)	1.434 (0.221)	1.538 (0.208)	1.441 (0.194)
Res. variance	1036.40	1036.42	1036.44	1036.46	1036.48	1036.50

Notes: The constructed variable is described in Section 6.4. The results are given for imputed values of the correlation $0, 0.2, \ldots, 1$.

$$\begin{aligned} \boldsymbol{\beta} &= (\mathbf{I} - \mathbf{u}_2^{-1}\boldsymbol{\varepsilon}_2 + \mathbf{u}_2^{-1}\boldsymbol{\varepsilon}_2\mathbf{u}_2^{-1}\boldsymbol{\varepsilon}_2 - \ldots)(\mathbf{u}_2^{-1}\mathbf{u}_1 + \mathbf{u}_2^{-1}\boldsymbol{\varepsilon}_1) \\ &\sim \mathbf{u}_2^{-1}\mathbf{u}_1 + \mathbf{u}_2^{-1}\boldsymbol{\varepsilon}_1 - \mathbf{u}_2^{-1}\boldsymbol{\varepsilon}_2\mathbf{u}_2^{-1}\mathbf{u}_1 - \mathbf{u}_2^{-1}\boldsymbol{\varepsilon}_2\mathbf{u}_2^{-1}\boldsymbol{\varepsilon}_1 , \end{aligned}$$

and so the mean and variance matrix of $\boldsymbol{\beta}$ are approximated as

$$\begin{aligned} \mathbf{E}(\tilde{\boldsymbol{\beta}}) &\sim \mathbf{u}_2^{-1}\mathbf{u}_1 - \mathbf{u}_2^{-1}\left\{\mathrm{tr}\left(\mathbf{u}_2^{-1}\boldsymbol{\Sigma}_{21.1}\right), \mathrm{tr}\left(\mathbf{u}_2^{-1}\boldsymbol{\Sigma}_{21.2}\right), \ldots, \mathrm{tr}\left(\mathbf{u}_2^{-1}\boldsymbol{\Sigma}_{21.p}\right)\right\} \\ \mathrm{var}(\tilde{\boldsymbol{\beta}}) &\sim \mathbf{u}_2^{-1}\left(\boldsymbol{\Sigma}_1 - \mathbf{A} - \mathbf{A}^\top + \mathbf{B}\right)\mathbf{u}_2^{-1} , \end{aligned}$$

where $\mathbf{A}$ is the $p \times p$ matrix with columns $\boldsymbol{\Sigma}_{21.h}\mathbf{u}_2^{-1}\mathbf{u}_1$, and $\mathbf{B}$ is the $p \times p$ matrix with elements $\mathrm{tr}\left(\mathbf{u}_2^{-1}\mathbf{u}_1\mathbf{u}_1^\top\mathbf{u}_2^{-1}\boldsymbol{\Sigma}_{2.hh'}\right)$, $1 \le h, h' \le p$.

For multilevel regression a number of data summaries, various within-cluster sums of squares and crossproducts, are required. An algorithm for maximizing the log-likelihood for the population can be used, with the population quantities replaced by their sample counterparts. The equations for such an algorithm involve complex functions of the summaries, and, as a consequence the delta method leads to unwieldy equations, in particular for estimation of variances and covariances. Nevertheless, the outlined approach can easily be applied without bias correction and derivation of the sampling variance matrix of the estimators. The impact of the unknown covariance or correlation matrix can be explored by imputing several extreme values of the matrix, such as the matrix of zeros or other singular matrices. The same approach can, in principle, be applied to structural model equations and factor analysis.

7 Two-stage clustered sampling design

This section derives the equations for the method of Potthoff *et al.* (1992) for the two-stage (three-level) clustered sampling design. It arises, for instance, when the replicate groups in the 1990 Math

Trial State Assessment data are associated with variation. The NAEP Assessment for USA employs a stratified probability clustered sample with two stages of clustering. This section extends the model-based methods to such sampling designs. We present first the details for the two-stage clustered design with no stratification. Incorporating stratification is relatively straightforward because it merely corresponds to collating independent information across the strata.

We use the term 'group' for the same aggregate sampling units (replicate groups) as above, even though these units are now associated with *random* variation. We consider the model

$$y_{ijk} = m + \alpha_k + \delta_{jk} + \varepsilon_{ijk}, \tag{52}$$

where $\{\alpha_k\}_k$, $\{\delta_{jk}\}_{jk}$, and $\{\varepsilon_{ijk}\}_i$ are $2 + N_2$ mutually independent random samples from centered distributions with respective variances ω^2, τ^2, and $\{\sigma^2_{jk}\}_{jk}$. These variances are referred to as *variance components*. The covariance of two observations in the same cluster is $\tau^2 + \omega^2$ and the covariance of two observations from the same group but different clusters is ω^2. Since these covariances can, in principle, be negative, it is meaningful, though counterintuitive, to consider negative 'variance' components, so long as the variance matrices for the sample and the target population are non-negative definite.

This model differs from that for the stratified clustered sampling only by the assumptions for the group deviations α_k. In Section 3.4 these deviations are assumed to be unknown constants (fixed), whereas here we assume them to be a set of i.i.d. random variables. As in Section 3.4 we do not assume specific distributions for the random variables in (52). We focus on the ratio estimator of the population mean for y:

$$\mu = \frac{\sum_k \sum_j \sum_i w_{ijk} y_{ijk}}{\sum_k \sum_j \sum_i w_{ijk}}.$$

We define the stage-level weighted sample means

$$\begin{aligned} \mu_{jk} &= \frac{\sum_i w_{A\,ijk} y_{ijk}}{n_{A\,jk}} \\ \mu_k &= \frac{\sum_j w_{B\,jk} \mu_{jk}}{n_{B\,k}}, \end{aligned}$$

where

$$\begin{aligned} w_{A\,ijk} &= \frac{W_{jk}}{\sum_i w^2_{ijk}} w_{ijk} \\ w_{B\,jk} &= \frac{W_k}{\sum_j W^2_{jk}} W_{jk}; \end{aligned}$$

$W_{jk} = \sum_i w_{ijk}$ and $W_k = \sum_j W_{jk}$, and $n_{A\,jk} = \sum_i w_{A\,ijk}$ and $n_{B\,k} = \sum_j w_{B\,jk}$ are the *effective sample sizes* of students within schools, and schools within 'strata'. Note also that $n_{A\,jk} = \sum_i w^2_{A\,ijk}$ and $n_{B\,k} = \sum_j w^2_{B\,jk}$. In analogous notation, the population weighted sample mean is expressed as

$$\mu = \frac{\sum_k w_{C\cdot k}\mu_k}{n_{C\cdot}},$$

where $w_{C\cdot k} = WW_k / \sum_j W_k^2$, $n_{C\cdot} = \sum_k w_{C\cdot k}^h$, for $h = 1, 2$, and $W = \sum_k W_k$.

Further, we denote $\mu_k = \mu + \alpha_k$ and $\mu_{jk} = \mu + \alpha_k + \delta_{jk}$, so that μ_{jk} and μ_k are estimators of the realized values of μ_{jk} and μ_k, respectively. These estimators are conditionally unbiased given μ_{jk} and μ_k, respectively.

The moment method of estimating the variance components ω^2, τ^2, and $\{\sigma_{jk}^2\}_{jk}$ is based on the following ANOVA-like (weighted) sums of squares:

$$\begin{aligned} v_{A\cdot jk} &= \frac{\sum_i w_{A\cdot ijk}(y_{ijk} - \mu_{jk})^2}{n_{A\cdot jk} - 1} \\ v_{B\cdot k} &= \sum_j u_{jk}(\mu_{jk} - \mu_k)^2 \\ v_{C\cdot} &= \sum_k u_k(\mu_k - \mu)^2. \end{aligned} \tag{53}$$

The sets of weights $\{u_{jk}\}$ and $\{u_k\}$ can be arbitrary non-negative numbers; their choice is discussed below. First, we evaluate the expectations of these statistics as (linear) functions of the variance components; then linearly combine them and solve the resulting moment equations by setting these statistics equal to their expectations.

For derivation of the expectations of the sum-of-squares statistics in (53) the following identities are useful:

$$\begin{aligned} \mu_{jk} - \mu_{jk} &= \frac{\sum_i w_{A\cdot ijk}\varepsilon_{ijk}}{n_{A\cdot jk}} \\ \mu_k - \mu_k &= \frac{\sum_j w_{B\cdot jk}\left(\delta_{jk} + \frac{\sum_i w_{A\cdot ijk}\varepsilon_{ijk}}{n_{A\cdot jk}}\right)}{n_{B\cdot k}} \\ \mu - \mu &= \frac{\sum_k w_{C\cdot k}\left(\alpha_k + \frac{\sum_j w_{B\cdot jk}\delta_{jk}}{n_{B\cdot k}}\right)}{n_{C\cdot}} + \frac{\sum_k w_{C\cdot k}\frac{\sum_j w_{B\cdot jk}\frac{\sum_i w_{A\cdot ijk}\varepsilon_{ijk}}{n_{A\cdot jk}}}{n_{B\cdot k}}}{n_{C\cdot}}. \end{aligned}$$

Hence

$$\begin{aligned} \mathbf{E}\{(\mu_{jk} - \mu_{jk})^2\} &= \frac{\sigma_{jk}^2}{n_{A\cdot jk}} \\ \mathbf{E}\{(\mu_k - \mu_k)^2\} &= \frac{\tau^2}{n_{B\cdot k}} + \frac{1}{n_{B\cdot k}^2}\sum_j \frac{w_{B\cdot jk}^2}{n_{A\cdot jk}}\sigma_{jk}^2 \\ \mathbf{E}\{(\mu - \mu)^2\} &= \frac{\omega^2}{n_{C\cdot}} + \frac{\tau^2}{n_{C\cdot}^2}\sum_k \frac{w_{C\cdot k}^2}{n_{B\cdot k}} + \frac{1}{n_{C\cdot}^2}\sum_k \frac{w_{C\cdot k}^2}{n_{B\cdot k}^2}\sum_j \frac{w_{B\cdot jk}}{n_{A\cdot jk}}\sigma_{jk}^2. \end{aligned} \tag{54}$$

Note that $\mathbf{E}\{(\mu_{jk} - \mu_{jk})^2\} = \mathrm{var}(\mu_{jk} \mid \mu_{jk})$ and $\mathbf{E}\{(\mu_k - \mu_k)^2\} = \mathrm{var}(\mu_k \mid \mu_k)$.

For the within-school weighted sum of squares we have

$$\begin{aligned}\mathbf{E}(v_{A\,jk}) &= \frac{1}{n_{A\,jk}-1}\mathbf{E}\Big\{\sum_i w_{A\,ijk}(y_{ijk}-\mu_{jk})^2 + n_{A\,jk}(\mu_{jk}-\mu_{jk})^2 \\ &\quad - 2(\mu_{jk}-\mu_{jk})\sum_i w_{A\,ijk}(y_{ijk}-\mu_{jk})\Big\} \\ &= \sigma^2_{jk},\end{aligned}$$

since $y_{ijk} - \mu_{jk} = \varepsilon_{ijk}$ and $\sum_i w_{A\,ijk}(y_{ijk} - \mu_{jk}) = n_{A\,jk}(\mu_{jk} - \mu_{jk})$. Thus, $v_{A\,jk}$ is an unbiased estimator of σ^2_{jk}.

Using the decomposition

$$\begin{aligned}v_{B\,k} &= \sum_j u_{jk}\Big\{(\mu_{jk}-\mu_{jk})^2 + (\mu_{jk}-\mu_k)^2 + (\mu_k-\mu_k)^2 \\ &\quad + 2(\mu_{jk}-\mu_{jk})(\mu_{jk}-\mu_k) - 2(\mu_{jk}-\mu_k)(\mu_k-\mu_k) \\ &\quad - 2(\mu_{jk}-\mu_{jk})(\mu_k-\mu_k)\Big\}\end{aligned}$$

we obtain

$$\begin{aligned}\mathbf{E}(v_{B\,k}) &= \frac{\tau^2}{n_{B\,k}}\sum_j u_{jk}(n_{B\,k} + 1 - 2w_{B\,jk}) \\ &\quad + \sum_j \sigma^2_{jk}\frac{u_{jk}}{n_{A\,jk}}\left(1 - 2\frac{w_{B\,jk}}{n_{B\,k}}\right) + \frac{1}{n^2_{B\,k}}\sum_j \sigma^2_{jk}\frac{w^2_{B\,jk}}{n_{A\,jk}}\sum_j u_{jk}\end{aligned} \tag{55}$$

When $\sum_j u_{jk} = n_{B\,k}$ (55) simplifies to

$$\begin{aligned}\mathbf{E}(v_{B\,k}) &= \tau^2\left(n_{B\,k} + 1 - \frac{2\sum_j u_{jk}w_{B\,jk}}{n_{B\,k}}\right) \\ &\quad + \frac{1}{n_{B\,k}}\sum_j \frac{\sigma^2_{jk}}{n_{A\,jk}}\left\{u_{jk}(n_{B\,k} - 2w_{B\,jk}) + w^2_{B\,jk}\right\}.\end{aligned} \tag{56}$$

Note that although $\sum_j u_{jk} = n_{B\,k}$ is a restrictive condition, because it imposes relative weights on the clusters within 'strata', the relative weighting of the 'strata' can be compensated in the linear combination of the group-level sums of squares $v_{B\,k}$. For $u_{jk} = w_{B\,jk}$ we have

$$\mathbf{E}(v_{B\,k}) = \tau^2(n_{B\,k} - 1) + \frac{1}{n_{B\,k}}\sum_j \sigma^2_{jk}\frac{w_{B\,jk}(n_{B\,k} - w_{B\,jk})}{n_{A\,jk}}. \tag{57}$$

If in addition the within-school totals of weights W_{jk} are constant within 'strata' then

$$\mathbf{E}(v_{B\,k}) = (n_k - 1)\left(\tau^2 + \frac{1}{n_k}\sum_j \frac{\sigma^2_{jk}}{n_{A\,jk}}\right).$$

BEST COPY AVAILABLE

where $n_k = n_{B\,k}$ is the school-sample size in group k. Note that the 'degrees of freedom' appear in this equation. The identities

$$\begin{aligned} \frac{w_{B\,jk}}{n_{B\,k}} &= \frac{W_{jk}}{W_k} \\ n_{A\,jk} &= \frac{\sum_i w_{ijk}^2}{W_{jk}^2} \\ n_{B\,k} &= \frac{\sum_j W_{jk}^2}{W_k^2} \end{aligned}$$

can be used to express (55) – (57) in terms of the sampling weights w_{ijk} and their stage-level totals W_{jk} and W_k.

By similar operations applied to the between-group sum of squares we obtain

$$\begin{aligned} \mathbf{E}(r_{C\cdot}) &= \frac{\omega^2}{n_{C\cdot}} \left\{ (n_{C\cdot}+1) \sum_k n_k - 2 \sum_k w_{C\cdot k} n_k \right\} \\ &\quad + \tau^2 \left(\sum_k \frac{n_k}{n_{B\,k}} - \frac{2}{n_{C\cdot}} \sum_k \frac{w_{C\cdot k}}{n_{B\,k}} n_k + \frac{1}{n_{C\cdot}^2} \sum_k \frac{w_{C\cdot k}^2}{n_{B\,k}} \sum_k n_k \right) \\ &\quad + \sum_k \frac{n_k}{n_{B\,k}^2} \left(1 - 2\frac{w_{C\cdot k}}{n_{C\cdot}} \right) \sum_j \frac{w_{B\,jk}^2}{n_{A\,jk}} \sigma_{jk}^2 \\ &\quad + \frac{1}{n_{C\cdot}^2} \sum_k n_k \sum_k \frac{w_{C\cdot k}^2}{n_{B\,k}^2} \sum_j \frac{w_{B\,jk}^2}{n_{A\,jk}} \sigma_{jk}^2 . \end{aligned} \tag{58}$$

We can assume, without loss of generality, that $\sum_k n_k = n_{C\cdot}$; then

$$\begin{aligned} \mathbf{E}(r_{C\cdot}) &= \omega^2 \left\{ (n_{C\cdot}+1) - \frac{2\sum_k w_{C\cdot k} n_k}{n_{C\cdot}} \right\} \\ &\quad + \frac{\tau^2}{n_{C\cdot}} \left[\sum_k \frac{1}{n_{B\,k}} \left\{ n_k (n_{C\cdot} - 2w_{C\cdot k}) + w_{C\cdot k}^2 \right\} \right] \\ &\quad + \frac{1}{n_{C\cdot}} \sum_k \left\{ \frac{n_k}{n_{B\,k}^2} (n_{C\cdot} - 2w_{C\cdot k}) + w_{C\cdot k}^2 \right\} \sum_j \frac{w_{B\,jk}^2}{n_{A\,jk}} \sigma_{jk}^2 . \end{aligned}$$

For $n_k = w_{C\cdot k}$ further simplification takes place:

$$\begin{aligned} \mathbf{E}(r_{C\cdot}) &= \omega^2 (n_{C\cdot} - 1) + \frac{\tau^2}{n_{C\cdot}} \sum_k \frac{w_{C\cdot k}(n_{C\cdot} - w_{C\cdot k})}{n_{B\,k}} \\ &\quad + \frac{1}{n_{C\cdot}} \sum_k \frac{w_{C\cdot k}(n_{C\cdot} - w_{C\cdot k})}{n_{B\,k}^2} \sum_j \frac{w_{B\,jk}}{n_{A\,jk}} \sigma_{jk}^2 . \end{aligned} \tag{59}$$

These expressions for $\mathbf{E}(r_{C\cdot})$ can be rewritten in terms of the sampling weights and their totals using the identities

$$\begin{aligned} \frac{w_{C\cdot k}}{n_{C\cdot}} &= \frac{W_k}{W} \\ n_{C\cdot} &= \frac{W^2}{\sum_k W_k^2} . \end{aligned}$$

For example, the coefficient of τ^2 in (59) is equal to

$$\frac{1}{\sum_k W_k^2}\left(W\sum_k \frac{\sum_j W_{jk}^2}{W_k} - \sum_k\sum_j W_{jk}^2\right).$$

When the group-level totals of weights are constant then (59) further simplifies to

$$\mathbf{E}(v_C) \;=\; (K-1)\left(\omega^2 + \frac{\tau^2}{K}\sum_k \frac{1}{n_{B\cdot k}} + \frac{1}{K}\sum_k \frac{1}{n_{B\cdot k}^2}\sum_j \frac{w_{B\cdot jk}}{n_{A\cdot jk}}\sigma_{jk}^2\right),$$

where K is the number of groups.

7.1 Estimators of the variance components

The within-school variances are estimated by $v_{A\cdot jk}$. If schools share within-school variances, the statistics $v_{A\cdot jk}$ can be combined to form an estimator of the common variance. For instance, if all schools within a group have a common within-school variance σ_k^2 then

$$\frac{1}{\sum_j n_{A\cdot jk} - 1}\sum_j (n_{A\cdot jk}-1)v_{A\cdot jk}$$

is an unbiased estimator of σ_k^2.

In general, we have

$$\begin{aligned}
\mathbf{E}(v_{B\cdot k}) &= D_{B\cdot k}\tau^2 + \sum_k\sum_j D_{B\cdot jk}\sigma_{jk}^2 \\
\mathbf{E}(v_C) &= D_C\omega^2 + D_{C2}\tau^2 + \sum_k\sum_j D_{C\cdot jk}\sigma_{jk}^2,
\end{aligned}$$

where the coefficients $D_{B\cdot k}$, $D_{B\cdot jk}$, D_{C2}, and $D_{C\cdot jk}$ are as in (55) and (58), or their special cases. Moment estimators for the school- and group-level variances are the solutions of the pair of linear equations

$$\begin{aligned}
\sum_k s_k v_{B\cdot k} &= \sum_k s_k D_{B\cdot k}\tau^2 + \sum_k s_k \sum_j \sigma_{jk}^2 \\
v_C &= D_C\omega^2 + \sum_k D_{C\cdot k}\tau^2 + \sum_k\sum_j D_{C\cdot jk}\sigma_{jk}^2,
\end{aligned} \tag{60}$$

where s_k are a set of non-negative constants;

$$\begin{aligned}
\tau^2 &= \frac{\sum_k s_k v_{B\cdot k} - \sum_k s_k \sum_j \sigma_{jk}^2}{\sum_k s_k D_{B\cdot k}} \\
\omega^2 &= \frac{v_C - \sum_k D_{C\cdot k}\tau^2 - \sum_k \sum_j D_{C\cdot jk}\sigma_{jk}^2}{D_C}.
\end{aligned} \tag{61}$$

The weights $w_{C\cdot k}$ appear to be natural choices for the coefficients s_k. The estimates of the variance components are then substituted for the true values in (54) to obtain an estimate of the sampling variance $\mathrm{var}(\mu)$

of the population mean estimator μ. Using different sets of weights for the moment matching equations, estimating some of the variance components by smoothing techniques involving multiple regression opens up a variety of possibilities which require further research.

7.2 Stratified two-stage clustered design

For purposes of statistical analysis the design of the NAEP surveys is described as a stratified two-stage weighted sampling design. Thus, within each group a small number, usually two, *primary sampling units* (PSU) is 'selected' (in reality, a different stratification is used, but after selecting the primary sampling units they are paired into replicate pairs of PSUs). Then clusters (schools, or 'consolidated' schools) are selected from each selected PSU, and finally students are sampled from each selected school.

Since within each group we have a two-stage clustered design, the estimators of the variance components and of the sampling variance of the mean carry over to stratified design, with suitable weights for combining the within-stratum estimates of the variance components.

8 Summary

The report describes a class of model-based methods for estimation of the population mean in stratified clustered sampling. Importance of the adjustment of weights is assessed by an approach considering the sampling variation of the adjusted weights and its (variance) components. The methods are non-iterative, and the resulting estimators are more efficient than the jackknife estimators for a variety of datasets obtained from the 1990 Math Trial State Assessment. The methods can be extended to two-stage clustering. A general method for estimation of more complex population summaries, such as regression coefficients, is outlined. It is based on the estimators of the population means, applied to various quadratic functions of the explanatory and outcome variables. There are no distributional assumptions in model-based methods, apart from normality of the sample means. Model-based methods use only the final adjusted weights; the replicate weights can be disposed of, thus radically reducing the size of the dataset and simplifying data handling procedures. The principal advantage of the model-based methods is in efficiency and small bias of the estimators of standard errors for the population mean. Contrary to theoretical claims, the (NAEP) operationally implemented jackknife estimator of the sampling variance is not unbiased.

9 Appendix. Data analysis with Splus

This section describes and documents functions and programs written in Splus for processing and analysis of 1990 Math Trial State Assessment data.

9.1 Data input

The data were obtained from the User Tapes, and in the process of transferring them to the Sun-Workstation on which the analysis would take place only a subset of 120 variables was selected. They included all the resampling weights, plausible values for the composite proficiency scores, information about student background, and a subset of responses to cognitive and attitudinal items. The file containing this dataset is named **NJ.dat** for New Jersey, and similarly for other states.

The data are input into the Splus environment by the following Splus expressions:

```
Ncols _ 120
NJdat _ matrix(scan("NJ.dat"),ncol=Ncols,byrow=T)
```

The function *scan* 'scans', or reads, the dataset and temporarily stores it in a vector; by default, two numbers are separated by one or several spaces, one or several carriage returns, or their combination. This default can easily be overruled. The function *matrix* with the argument *byrow=T* 'shapes' this vector into a matrix with *Ncols* columns by filling up its elements row by row. There are reasonable defaults if the vector has a length which is not a multiple of *Ncols*, and the user is informed about it by a *warning*.

The number of students in the data is ascertained as the number of rows of the matrix *NJdat*:

```
Nstud _ dim(NJdat)[1]
```

Next, we sort the students by the schools and by strata. The scalars *Njack* and *Npair* are the indices (column numbers) for the stratum and the school within the stratum. Knowing that there are at most three schools within a stratum, the students can be sorted on the variable

$$3 \times \text{stratum number} + \text{school number within stratum}\,.$$

First, to ease the burden of typing complex expressions we define an Splus function *cls*:

```
cls _ function(c1,c2)
NJdat[,c1]%*%matrix(c2)
```

This function has two *arguments*; *c1* should be a vector, a list of column indices of *NJdat*, and *c2* a vector of the same length as *c1*. The function returns the linear combination of the columns *c1* of *NJdat* with coefficients *c2*. The vector *c2* has to be reshaped into a matrix because Splus distinguishes between vectors

and matrices with a single column or row. The value of the function *cls* is a vector even though it is generated by a matrix operation.

Sorting of the data is accomplished by the following expressions:

```
#  Npair and Njack are the column indices
#  for the replicate and jackknife

Npair _ 12
Njack _ 11

# sort records by schools

NJdat _ NJdat[sort.list(cls(c(Npair,Njack),c(1,3))),]
```

The hash '#' causes the rest of the line to be regarded as a comment. The function *sort.list* returns the permutation that would sort its argument in ascending order. This permutation is then applied to the rows of the matrix *NJdat*. The selection of columns of *NJdat* can be affected by an expression between the last comma above and the closing bracket ']'. No expression behind the comma is interpreted as 'all columns'.

For convenience, it is useful to calculate the delimiters of the schools in the data. We index the schools by integers $1, 2, \ldots$, where school 1 is represented by records $1, 2, \ldots, n_1$, school 2 by $n_1 + 1, \ldots, n_1 + n_2$, and so on. We refer to n_j as the school sample sizes.

```
Bot _ seq(Nstud)[!duplicated(cls(c(Npair,Njack),c(1,3)))]
Top _ c(Bot[-1]-1,Nstud)
Cnt _ Top - Bot + 1
Nclu _ length(Cnt)
```

In this sequence of expressions *Bot* is assigned all the indices (components of the vector $1, 2, \ldots, Nstud$), for which any value of the linear combination of replication and jackknife indices occurs for the first time. The exclamation mark '!' stands for negation, and the function *duplicated* returns a logical vector (vector of T's and F's) indicating whether the value of the component of the argument is equal to that of a previous component. Thus, *Bot* contains the indices of the first students from each school, *Top* is set to the indices of the last students from each school, *Cnt* to the number of students from each school, and *Nclu* to the number of schools in the dataset (*length* returns the number of components).

9.2 A data summary

The following expressions give the selected variable names and tabulate the categorical variables and compute quantiles and standard deviations for the quantitative variables. Note how the function *paste* is used for generating a set of similar names (character strings).

```
#  Tabulating the NJdat data
#  Tabulation

Vnames _ c("Sex","School","Race","IEP","LEP",
   "D.Sex","D.race","Urb.Stratum","Min.Stratum","Inc.Stratum",
   "Rep.Grp.1","Drop.Grp","JK.Fac","Weight",
    paste("SRWT",seq(1,56)),
   "Orig.WT","Num.Cor","Par.Ed.","Single.P","Sch.Math",
   "Perc.Math","T.Certf","T.Und.Mj","T.Grp.Mj","T.Math.Crs",
   "T.Emph.No.","T.Emph.PS","S.Policy","Problems","B003501A",
   "B003601A","B000901A","B000903A","B000904A","B000905A",
    paste("M",10100+seq(1,8),"B"),
   "M810201B", paste("M",10300+seq(1,3),"B"),
    paste("MPRCMP",seq(1,5)),
    paste("T023",c(201,301,302,311,312,307,308,313,401,402,
    411,412,407)))

length(Vnames)

quanti _ c(1,2,3,seq(13,72),75,seq(103,107))
categ _ seq(1,120)[-quanti]

NJTAB _ list()
for (i in categ)
NJTAB[[i]] _ TBL(i)

for (i in quanti)
NJTAB[[i]] _ c(mean(NJdat[,i]),quantile(NJdat[,i],
c(0,.1,.25,.5,.75,.9,1)), sqrt(var(NJdat[,i])))

"Done"
```

The content of the list *NJTAB* is, naturally quite extensive, and therefore we reproduce only a small section of it, giving the summary for a categorical and a quantitative variable.

```
.
.
.
$"Sch.Math , No.  75":
```

```
   Mean Minimum    10%  25% Median   75%     90% Maximum St. Dev.
 3371.8  -25021 -14404 -269   4920 11410 14975.5   19804  10445.4

$"Perc.Math , No.  76":
   1    2   3  9
 725 1392 555 38
  .
  .
  .
```

Variable No. 75 is the *School level math mean logit score* (multiplied by 10 000), No. 189 on the User Tapes, and No. 76 is *Student's perception of mathematics*, No. 190 on the User Tapes.

9.3 Jackknife

In this section we present an Splus function for jackknife estimation of subpopulation means. Specifically, we consider estimation of the mean proficiency for a subpopulation given by a condition described in terms of the variables in the dataset. The vector of values of this (logical) variable on the sampled students is the argument of the Splus function *Jackf*. The default argument is T, that is, the entire population.

The function starts with extracting the dataset for the subsample corresponding to the subpopulation (*NJd*), and the sample size of this subsample (*nStu*). The *permanent* assignment symbol '<< −' has the effect of its left-hand side to be written in the directory available at the entry into Splus. Other objects created within the function are *temporary*; with the exception of the last expression of the function they are not available after the function is successfully evaluated. If an error occurs during evaluation, the directory remains intact.

The objects *bot*, *top* and *cnt* are the analogues of the vectors *Bot*, *Top* and *Cnt* for the subsample. The object *Twt* (totals of the weights) is a vector of length *Njr* (number of strata + 1), and its components are the total of adjusted weights (the first component), and the totals of the jackknife replicate weights for each pseudoanalysis. *Cwt* is the index of the adjusted weights, and the adjusted weights are followed by the jackknife replicate weights in each record of the dataset.

The permanent assignment of *Twt*, as well as of *NJd*, is essential because these objects are used in another function: the function *jkmean* evaluates the sample mean or the jackknife pseudo-means for a set of plausible values. The indices for the analysis (sample or jackknife) and for the plausible value are encoded in the argument j. The function *jkmean* is used via the *apply* function to create the vector of these means. The function *apply* has three arguments; an array A, an integer i, and a function f. The function f is applied to each subplane of the i-th dimension of A. In our case, A is a column vector of integers $0, 1, \ldots, Njr \times Npr - 1$ (*Npr* is the number of plausible values, equal to 5). For a given integer the associated pseudo-analysis and the plausible value are given by integer division (%/%) and remainder (%%), respectively; see the declaration of the function *jkmean*. The values returned by the *apply* function

BEST COPY AVAILABLE

using *jkmean* are reshaped into a $Njr \times Npr$ matrix; it contains the sample and pseudo-means (rows) for the sets of plausible values (columns).

The function *ssq* collects the results of the Npr jackknife analyses: the sample (weighted) means, the jackknife estimates of the means, and the jackknife estimates of the sampling variances. Since the decimal placing was ignored at the input, division by 100 and 10 000 brings the results onto the appropriate scale. The function is then applied to each column of *sbar*, that is, for each plausible value, thus generating a $3 \times Npr$ matrix.

The matrix *svars* is augmented by the means across the plausible values and its third row (sampling variance) is augmented by the observed variance of the sampling variances. Finally, the variances are transformed to standard errors (deviations), labels are attached to the rows and columns of *svars*, and the function returns a list containing:

- numbers of students and schools;
- counts of students within schools;
- the matrix *svars*;
- user, system and elapsed times for evaluation of the function (note that the value of *start* is assigned by the first expression of the function).

The expressions constituting the function *jkmean* are enclosed in braces { }; for functions containing a single expression, such as *jkmean* and *ssq*, the braces are redundant.

The following is an example of using the function *jkmean*. The jackknife analysis is performed for the subpopulation of all students who (would have) responded 'A' ('strongly agree') to the question about student's perception of mathematics (variable No. 190 in the user tape).

```
JKre762 _ Jackf(NJdat[,76]==1)
JKre762
```

The first expression is an assignment; the second, quoting the name of the object, displays the object given below:

```
$Students.clusters:
[1] 725 104


$counts:
  [1]  8  4  4 10  6 12  7 10  5  3  9  7 13  4  6  6 11  6  1 14
 [21]  9  8 11  5  3 14 12  4  2  6  8 12 11 12  4  6  5  5  6 10
 [41]  6  7  2  7 13  7 13  5  9  3  5  4  5  4  2  3 10 11  8  3
 [61] 11  3  6  6  5  4 14  6  5  7  2  8  6  6 10 11  6  8  7 10
 [81]  7 11  7 12  8  5  5  2  2  4  4  5  8 15  9  6  5  5  9  5
```

```
[101]  4  8  3  9
```

```
$estimates:
             P.vl. 1  P.vl. 2  P.vl. 3  P.vl. 4  P.vl. 5   Ov-ll
Weight. mean 279.837  279.592  279.516  279.910  279.732  279.717
    JK. mean 279.828  279.579  279.501  279.905  279.718  279.706
 JK. st.err.   1.681    1.707    1.722    1.659    1.727    1.707
```

Thus, the subsample contains 725 students from 104 schools (each school is represented in the subsample). The mean proficiency is 279.71 – the difference between the jackknife and ratio (weighted) estimates is trivial. The estimated standard error of the jackknife estimator is 1.71.

```
#  Jackknife analysis for subsets
#  cases selected by   seval

#  1990 New Jersey state trial assessment

#  The function requires only the dataset  NJdat
#  and the logical vector
#  for the selected subset,  e.g.  Jackk(NJdat[,74]==2)
#  (!!)  or the vector of subscripts

#  Constants to be set
#  Cpr ...  the first column of plausible values
#     Cpr = 103
#  lp  ...  number of Plausible Values
#           lp = 5,  and  Ip=seq(1,lp)
#  Cwt ...  the column of weights (followed by the JK weights)
#           Cwt = 14
#  Njr ...  number of PSUs
#           Njr = 57
#  The default is the entire dataset  (2710 students)



Jackf _ function(seval = T)
{
```

```
start _ proc.time()

#  select the students

NJd <<- NJdat[seval,  ]
nStu <- dim(NJd)[1]

#  the delimiters for the clusters

bot <- seq(1, nStu)[!duplicated(NJd[, c(Npair, Njack)] %*%
matrix(c(1, 3)))]
top <- c(bot[-1] - 1, nStu)
cnt <- top - bot + 1

#  the total weights

Twt <<- apply(NJd[, Cwt:(Cwt + Njr - 1)], 2, sum)

#  analysis for each replicate and plausible value

jkmean <- function(j)
{
jj <- j %% (Njr)
sum(NJd[, Cpr + j %/% Njr] * NJd[, Cwt + jj])/Twt[jj + 1]
}

#  Jackknife means

sbar <- matrix(apply(matrix(seq(0, Npr * Njr - 1)), 1, jkmean),
Njr, Npr)

#  jackknife results (means and variances) for each pl. value

ssq <- function(yb)
c(yb[1]/100, mean(yb[-1])/100, sum((yb[-1] - yb[1])^2)/10000)

svars <- apply(sbar, 2, ssq)
```

```
#  variance --- standard error

svars <- cbind(svars, apply(svars, 1, mean))
svars[3, lp + 1] <- svars[3, lp + 1] + 1.2 * var(svars[2,  ])
svars[3,  ] <- sqrt(svars[3,  ])

dimnames(svars) <- list(c("Weighted mean", "JK. mean",
"JK. st.err."), c(paste("P.vl.", Ip), "Ov-ll"))

#  results

list(Students.clusters = c(nStu, length(cnt)), counts = cnt,
estimates = svars, proc.time=proc.time()-start)  }
```

9.4 Model based estimation

The Splus function *VCmg* listed below executes the set of model-based estimation procedures described in Section 3.4.

```
#  Function for model based estimation of NAEP data subsets
#  Filename NS.md   (started 6/8/92)   ---   method based on
#  Potthoff et al.  JASA, 1992

#  cases selected by   seval

#  1990 New Jersey state trial assessment

#  The function requires only the dataset  NJdat
#      and the logical vector
#  for the selected subset,  e.g.  Jackk(NJdat[,74]==2)
#  (!!)  or the vector of subscripts

#  Use function  TBL  to tabulate a NAEP variable

#  Constants to be set
#  Cpr ...  the first column of plausible values
#           Cpr = 103
#  lp  ...  number of Plausible Values
#           lp = 5,  and  Ip=seq(1,lp)
```

```
#  Cwt ...  the column of weights
#           Cwt = 14
#  Njr ...  number of PSUs
#           Njr = 57
#  The default is the entire dataset  (2710 students)

#  STR ...  column of the stratifying variable,  default ...
#           STR = 11
#  other option ...   STR = 9 (4   strata)

VCmg <- function(seval = T,STR=11)
{
start <- proc.time()
#  select the students

y <- NJdat[seval,c(Npair,Njack,STR,Cwt,Cpr+Ip-1)]

str <- y[,3]
nStu <- dim(y)[1]

#  the delimiters for the clusters

bot <- seq(1, nStu)[!duplicated(y[, c(1,2)] %*%
matrix(c(1, 3)))]
top <- c(bot[-1] - 1, nStu)
cnt <- top - bot + 1
Ncl <- length(cnt)
clu <- rep(seq(1,Ncl),cnt)

  w <- y[,4]/1000
y <- y[,4+Ip]/100

#  stratifying variable (student- and cluster-level)
#  recode to strata 1,2, ..., nstr

cstr <- unique(str)
str <- match(str,cstr)
Str <- str[top]
```

```
nstr <- length(cstr)

#  the delimiters for the strata

Sbot <- seq(1,Ncl)[!duplicated(Str)]
Stop <- c(Sbot[-1]-1,Ncl)
Scnt <- Stop-Sbot+1

#  total weight

W <- sum(w)

#  sample means

Wmn <- t(y)%*%matrix(w)/W

#  within-cluster means

tcls <- cbind(tapply(w,clu,sum), tapply(w^2,clu,sum))
for (i in Ip)
tcls <- cbind(tcls, tapply(y[,i]*w, clu, sum)/tcls[,1])

#   normalized weights and effective sample sizes  A
#   estimate of the within-cluster variance

nA <- tcls[,1]^2/tcls[,2]

S2w <<- matrix(0,Ncl,lp)
for (i in Ip)
S2w[,i] <<- tapply((y[,i]-tcls[clu,2+i])^2*w, clu, sum) *
tcls[,1]/tcls[,2] / (nA-1)

  S2W <- apply(S2w[nA>1,],2,mean)
S2w[nA==1,] <<- S2W

#  within-stratum totals, only when more than 2 PSUs

Snb <- S2w/matrix(nA,nrow=length(nA),ncol=lp)
```

```
Snb <- Snb[Sbot,]+Snb[Stop,]
SnA <- nA[Sbot]+nA[Stop]
Swt <- tcls[Sbot,1]+tcls[Stop,1]

#  between-cluster sums of squares

difs _ (tcls[Sbot,-c(1,2)]-tcls[Stop,-c(1,2)])^2

v2I <-  t(difs)%*%matrix(Swt)
v2II <- t(difs)%*%matrix(SnA)
v2III <- apply(difs/Snb,2,sum)

#  between-variance estimates

BI <- (v2I - t(Snb[Sbot!=Stop,])%*%matrix(
Swt[Sbot!=Stop]))/sum(Swt[Sbot!=Stop])/2
BII <- (v2II - t(Snb[Sbot!=Stop,])%*%
matrix(nA[Stop]+nA[Sbot])[Sbot!=Stop,])/
sum(nA[Sbot!=Stop])/2
BIII <- (v2III - sum(Stop!=Sbot))/apply(1/Snb[Stop!=Sbot,],
2,sum)/2

#   reestimate

Swr <-  1/(matrix(2*v2I,nrow=length(Sbot),ncol=lp,byrow=T) +
                Snb)
v2R <- apply(difs*Swr,2,sum)
BR <- (v2R - apply((Swr*Snb)[Sbot!=Stop,],2,sum))/
apply(Swr[Sbot!=Stop,],2,sum)/2


#  within stratum totals

stra <- rep(seq(1,length(Scnt)),Scnt)
ucls <- cbind(tapply(tcls[,1],stra,sum),
tapply(tcls[,1]^2,stra,sum))

for (i in Ip)
```

```
ucls <- cbind(ucls,tapply(tcls[,1]*tcls[,2+i],stra,sum)/
        ucls[,1])

difs <- (tcls[,2+Ip]-ucls[stra,2+Ip])^2

#  estimates based on W_jk and nA

Excl _ -Sbot[Stop==Sbot]

Wcom <- tcls[,1] - tcls[,1]^2/ucls[stra,1]
vbI <- t(difs) %*% matrix(tcls[,1])
CI <- (vbI - t(S2w)%*%matrix(Wcom/nA))/sum(Wcom)

Wcom <- 1 - 2*tcls[,1]/ucls[stra,1]+ucls[stra,2]/ucls[stra,1]^2
vbII <- t(difs) %*% matrix(nA)
CII <- (vbII - t(S2w)%*%matrix(Wcom))/sum(Wcom*nA)

vbIII <- t(difs[Excl,])%*%matrix(nA[Excl]/Wcom[Excl])
CIII <- (vbIII - apply(S2w[Excl,],2,sum))/sum(nA[Excl])

#  reestimate

Snb _ 1/(matrix(vbI,nrow=length(nA),ncol=lp,byrow=T) +
S2w/matrix(nA,nrow=length(nA),ncol=lp))

vbR <- t((difs*Snb)[Excl,])%*%matrix(1/Wcom[Excl])

CR <- (vbR - apply((S2w/matrix(nA,nrow=length(nA),ncol=lp)*
Snb)[Excl,],2,sum))/apply(Snb[Excl,],2,sum)

#  variance of the weighted mean

nB <- W^2/sum(tcls[,1]^2)

varI <- 1/nB*(BI  + t(S2w)%*%matrix(tcls[,2])/
sum(tcls[,1]^2))
varII <- 1/nB*(BII + t(S2w)%*%matrix(tcls[,2])/
sum(tcls[,1]^2))
```

```
varIII <- 1/nB*(BIII + t(S2w)%*%matrix(tcls[,2])/
sum(tcls[,1]^2))
varR <- 1/nB*(BR  + t(S2w)%*%matrix(tcls[,2])/
sum(tcls[,1]^2))
varCI <- 1/nB*(CI  + t(S2w)%*%matrix(tcls[,2])/
sum(tcls[,1]^2))
varCII <- 1/nB*(CII  + t(S2w)%*%matrix(tcls[,2])/
sum(tcls[,1]^2))
varCIII <- 1/nB*(CIII  + t(S2w)%*%matrix(tcls[,2])/
sum(tcls[,1]^2))
varCR <- 1/nB*(CR  + t(S2w)%*%matrix(tcls[,2])/
sum(tcls[,1]^2))

#  results

indx <- c(2,3,4,5,6,7,8,9)

resm <- matrix(c(Wmn,varI,varII,varIII,varR,varCI,
varCII, varCIII, varCR, S2W, BI, BII, BIII, BR,
CI, CII, CIII, CR),  ncol=lp,byrow=T)
resm <- cbind(resm,apply(resm,1,mean))
resm[indx,lp+1] <- resm[indx,lp+1]+(1+1/lp)*
apply(resm[indx,Ip],1,var)
resm[indx,] <- sqrt(resm[indx,])

dimnames(resm) <- list(
c("Weighted mean", "SE I", "SE II", "SE III",
  "SE R", "SE CI", "SE CII", "SE CIII", "SE CR",
  "Within variance", "BV I", "BV II", "BV III",
  "BV R", "BV CI", "BV CII", "BV CIII","BV CR"),
  c(paste("Pl.val.", Ip), "Overall"))


list(Students.schools.strata = c(nStu, Ncl, nstr),
school.sizes = cnt, strata=cstr, PSUs.in.strata=Scnt,
Eff.size = nB, estimates = resm,
proc.time=proc.time()-start)
}
```

The arguments of the function are the subsample (given by a logical variable, with the entire sample as the default), and the stratification variable (default variable No. 11, variable No. 50 on the User Tapes). Only the minimal set of variables is selected into the array *y*. Then the decimal places for the weights and plausible values are adjusted. The object *lp* is the vector (1, 2, 3, 4, 5), defined as *seq(lp)*, where *lp* is the number of plausible values, equal to 5.

The delimiters for the clusters are set as in *jkmean* but, additionally, similar delimiters are set for the strata (with respect to clusters). Some care is necessary because clusters and even whole strata may absent in the subsample.

The array *tcls* contains the within-cluster totals of weights, squared weights, and weighted totals of plausible values; *nA* are the within-cluster effective sample sizes, and *S2W* is the matrix of estimates of the within-cluster variances σ_W^2. The notation for the sum of squares and variance estimator is similar to that in the text. The method based on pairs of PSU's uses the objects *Swa*, *Swt*, *v2*, *v2I*, *v2II*, and so on.

The other two methods require within-stratum totals of weights and means of plausible values. These are collected in the matrix *ncls*. The objects *vb*, *vbI*, *vbII*, and so on, are the counterparts of *v2* and *v2I*, *v2II* for the two methods. The sets of five (*lp*) estimates are then summarized in the last column of the matrix *resm*. The output of the function is a list containing the clustering structure of the analyzed subsample, the estimates, and information about processing time.

For illustration we give the analysis for response 'A' to the item *Student's perception of mathematics*.

```
$Students.schools.strata:
[1] 725 104  53

$school.sizes:
  [1]  8  4  4 10  6 12  7 10  5  3  9  7 13  4  6  6 11  6  1 14
 [21]  9  8 11  5  3 14 12  4  2  6  8 12 11 12  4  6  5  5  6 10
 [41]  6  7  2  7 13  7 13  5  9  3  5  4  5  4  2  3 10 11  8  3
 [61] 11  3  6  6  5  4 14  6  5  7  2  8  6  6 10 11  6  8  7 10
 [81]  7 11  7 12  8  5  5  2  2  4  4  5  8 15  9  6  5  5  9  5
[101]  4  8  3  9

$strata:
 6 51 102 149 196 244 295 359 410 476 492 541 600 652 677 732 792
 1  2   3   4   5   6   7   8   9  10  11  12  13  14  15  16  17

 813 868 926 968 989 1045 1101 1147 1195 1273 1301 1347 1404 1445
  18  19  20  21  22   23   24   25   26   28   29   30   31   32

 1498 1551 1598 1645 1725 1768 1809 1859 1915 1963 2013 2067 2125
   33   34   35   36   37   38   39   40   41   42   43   44   45

 2182 2253 2283 2337 2388 2475 2523 2587 2637
   46   47   48   49   50   51   52   53   54


$PSUs.in.strata:
 [1] 2 2 2 2 2 2 2 2 2 1 2 2 2 1 2 2 1 2 2 2 1 2 2 2 2 3 2 2 2 2
[31] 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 3

$Eff.size:
[1] 74.536

$estimates:
             P.vl. 1  P.vl. 2  P.vl. 3  P.vl. 4  P.vl. 5   Ov-ll
Weight. mean 279.837  279.592  279.516  279.910  279.732  279.717
Wtd. st.err.   1.537    1.572    1.525    1.440    1.536    1.531
Between-var. 102.215  112.575  105.488   81.197  104.311  101.157
```

```
 W.st.err II   1.497    1.553    1.515    1.428    1.520    1.511
Btwn-var. II  93.294  108.237  103.124   78.695  100.657   96.801
 W.st.err. R   1.716    1.694    1.639    1.530    1.665    1.670
 Btwn-var. R 145.696  142.324  132.325  101.063  135.083  131.298
 Within var. 643.821  614.549  613.602  679.647  623.104  634.944

$proc.time:
[1] 36.416  4.517 48.000  0.000  0.000
```

Note that the effective sample size is 74.5, much smaller than the number of clusters, 104. The first two estimates of the standard error are nearly identical (1.53 and 1.51), but differ appreciably from the third one (1.67) which is close to its jackknife counterpart (1.71).

For the entire sample there is a much closer agreement. The jackknife estimate of the standard error is 1.05, while its model-based counterparts are 1.09, 1.10, and 1.085.

9.5 Regression

In this section the Splus functions for fitting regression by the jackknife and model-based methods are given. The vector *rsel* identifies the students with complete records (for each variable using in constructing the explanatory variable), and *xvar* is the constructed explanatory variable. The function *JKreg* returns the results of a pseudoanalysis. These results are stored in *sreg* and are returned in a suitable format in the list *JKfit*.

```
#  An example with regression
#  Select all the subjects

rsel _ (NJdat[,91]<6)&(NJdat[,92]<6)&(NJdat[,93]<6)&
       (NJdat[,94]<6)&(NJdat[,95]<6)&(NJdat[,96]<6)&
       (NJdat[,97]<6)&(NJdat[,98]<6)

xvar _ NJdat[,91]+NJdat[,92]+NJdat[,93]+NJdat[,94]+
       NJdat[,95]+NJdat[,96]+NJdat[,97]+NJdat[,98]


JKreg _ function(j)
{
jw _ j%%Njr
jy _ j%/%Njr
```


BEST COPY AVAILABLE

```
#  regression of pl.value  jy  with weights  jw

ft _ lsfit(xv,NJd[,Njr+3+jy],NJd[,3+jw])$coef
matrix(c(ft,sum(NJd[,3+jw]*(NJd[,Njr+3+jy]-ft[1]-ft[2]*xv)^2)/
sum(NJd[,3+jw])))
}

ssq _ function(yb)
c(yb[1],mean(yb[-1]),sum((yb[-1]-yb[1])^2))


start _ proc.time()

   xv _ xvar[rsel]
   NJd _ NJdat[rsel,c(Npair,Njack,Cwt+seq(Njr)-1,Cpr+Ip-1)]

   NJd[,seq(Njr)+2] _ NJd[,seq(Njr)+2]/1000
   NJd[,Njr+2+Ip] _ NJd[,Njr+2+Ip]/100


   bot _ seq(1,dim(NJd)[1])[!duplicated(NJd[,c(1,2)]%*%
matrix(c(1,3)))]
   top _ c(bot[-1]-1, dim(NJd)[1])
   cnt _ top-bot+1
   clu _ rep(seq(1,length(cnt)), cnt)

   sreg _ apply(matrix(seq(Npr*Njr)-1),1,JKreg)

   sregi _ matrix(sreg[1,],Njr,Npr)
   sregs _ matrix(sreg[2,],Njr,Npr)
   sregv _ matrix(sreg[3,],Njr,Npr)

sresi _ apply(sregi,2,ssq)
sress _ apply(sregs,2,ssq)
sresv _ apply(sregv,2,ssq)

svari _ cbind(sresi,apply(sresi,1,mean))
svars _ cbind(sress,apply(sress,1,mean))
```

```
svarv _ cbind(sresv,apply(sresv,1,mean))

svari[3,lp+1] _ svari[3,lp+1] + (lp+1)/lp*var(svari[2,])
svars[3,lp+1] _ svars[3,lp+1] + (lp+1)/lp*var(svars[2,])

svars[3,] _ sqrt(svars[3,])
svari[3,] _ sqrt(svari[3,])
svarv _ svarv[-3,]

dimnames(svari) _ list(c("Intercept", "Jackknife intercept",
  "JK. st. err."),c(paste("Pl.val.",Ip),"Overall"))
dimnames(svars) _ list(c("Slope", "Jackknife slope",
  "JK. st. err."),c(paste("Pl.val.",Ip),"Overall"))
dimnames(svarv) _ list(c("Res. var.", "Jackknife res. var."),
c(paste("Pl.val.",Ip),"Overall"))

JKfit _ list(Students.clusters=c(dim(NJd)[1],length(cnt)),
  counts=cnt, intercept=svari, slope=svars, Res.variance=svarv,
proc.time=(proc.time()-start)[1:3])

JKfit
```

The program for the model-based estimator is given below. The nesting structure (given by *bot*, *top*, and *clu* for clusters, and *sbot*, *stpp* and *scnt* for strata) is required, but only one set of sampling weights is used. The notation is similar to that in other Splus functions and in the text.

```
#  Regression with NAEP State data using Model-based methods

#  Filename  NW.Reg

#  the same data as in NW.reg (jackknife)

#  Select all the subjects

rsel _ (NJdat[,91]<6)&(NJdat[,92]<6)&(NJdat[,93]<6)&
       (NJdat[,94]<6)&(NJdat[,95]<6)&(NJdat[,96]<6)&
       (NJdat[,97]<6)&(NJdat[,98]<6)

xvar _ NJdat[,91]+NJdat[,92]+NJdat[,93]+NJdat[,94]+
```

```
        NJdat[,95]+NJdat[,96]+NJdat[,97]+NJdat[,98]

#   the stratum indicator

STR _ 11

start _ proc.time()

###   xv _ xvar[rsel]

NJd _ NJdat[rsel,c(Npair,Njack,STR,Cwt,Cpr+Ip-1)]

NJd[,4] _ NJd[,4]/1000
NJd[,4+Ip] _ NJd[,4+Ip]/100

#   clustering

bot _ seq(1,dim(NJd)[1])[!duplicated(NJd[,c(1,2)]%*%
        matrix(c(1,3)))]
top _ c(bot[-1]-1, dim(NJd)[1])
cnt _ top-bot+1

Ncl <- length(cnt)
clu <- rep(seq(1,Ncl),cnt)

#    stratification

#  stratifying variable (student- and cluster-level)

str <- NJd[,3]
cstr <- unique(str)

#  recode to strata 1,2, ..., nstr

str <- match(str,cstr)
Str <- str[top]
nstr <- length(cstr)
```


90

```
#  the delimiters for the strata

sbot <- seq(1,Ncl)[!duplicated(Str)]
stpp <- c(sbot[-1]-1,Ncl)
scnt <- stpp-sbot+1

#  the  y-variate (first plausible value) and weights

w  <- NJd[,4]

#  total weight

Wt <- sum(w)

#  weighted within-cluster totals for 1, x, y, x^2, xy, y^2

Mfit _ list()
for (i in Ip)
{
yv <- NJd[,4+i]

Tcls _ cbind(tapply(w,clu,sum),tapply(xv*w,clu,sum),
  tapply(yv*w,clu,sum), tapply(xv^2*w,clu,sum),
  tapply(xv*yv*w,clu,sum),tapply(yv^2*w,clu,sum))

#  sample means of x, y, x^2, xy, y^2

WMn <- apply(Tcls,2,sum)/Wt

#  regression estimate

nume _ WMn[5] - WMn[2]*WMn[3]
deno _ WMn[4] - WMn[2]^2
beta _ nume/deno
alph _ WMn[3] - WMn[2]

sigm _ WMn[6] - WMn[3]^2 - WMn[5] + WMn[2]*WMn[3]
```

```
strr _ sqrt(sigm/deno/length(yv))


##  sampling variance estimation

dt _ cbind(xv,yv,xv^2,xv*yv,yv^2) - matrix(WMn[-1],
nrow=length(xv),ncol=5,byrow=T)

##  effective sample size (A)

W2s _ tapply(w^2,clu,sum)
nA _ Tcls[,1]^2/W2s

S2w <- array(0,c(Ncl,lp,lp))

for (i in 1:Ncl)
S2w[i,,] <- t(dt[bot[i]:top[i],])%*%dt[bot[i]:top[i],]/
(nA[i]-1)

##  within-stratum totals

stra <- rep(seq(length(scnt)),scnt)
Ucls <- matrix(tapply(Tcls[,1],stra,sum))

for (i in 2:dim(Tcls)[2])
Ucls <- cbind(Ucls,tapply(Tcls[,i],stra,sum)/Ucls[,1])

Tcls[,-1] _ Tcls[,-1]/matrix(Tcls[,1],nrow=dim(Tcls)[1],
ncol=dim(Tcls)[2]-1)

VrEst <- t(Tcls[,-1]-Ucls[stra,-1])%*%((Tcls[,-1]-Ucls[stra,-1])*
matrix(Tcls[,1], nrow=dim(Tcls)[1],ncol=dim(Tcls)[2]-1))


  WCom _ Tcls[,1]-Tcls[,1]^2/Ucls[stra,1]

  sMM _ S2w[1,,]*WCom[1]/nA[1]
  for (i in 2:Ncl)
```

```
  sMM _ sMM + S2w[i,,]*WCom[i]/nA[i]

  VHat2 _ (VrEst - sMM)/sum(WCom)

#  variance of the weighted mean

nB _ Wt^2/sum(Tcls[,1]^2)
varm _ S2w[1,,]*W2s[1]
for (i in 2:Ncl)
varm _ varm + S2w[i,,]*W2s[i]

varm _ (VHat2+varm/sum(Tcls[,1]^2))/nB

##   varm (5x5)   contains the variance matrix for
##   (x,y,x^2,xy,y^2)

##  sampling variation of the regression parameter estimate

##  numerator variance

nuvar _ varm[4,4] + varm[1,1]*varm[2,2] + varm[1,2]^2 +
WMn[2]^2*varm[2,2] + WMn[3]^2*varm[1,1] - 2*WMn[2]*varm[2,4] -
2*WMn[3]*varm[1,4] + 2*WMn[2]*WMn[3]*varm[1,2]

##  denominator variance

devar _ varm[3,3] + 2*varm[1,1]^2 + 4*WMn[2]^2*varm[1,1] -
4*WMn[2]*varm[1,3]

#  the estimated covariance of the numerator and denominator

covr _ varm[3,4] - 2*WMn[2]*varm[1,4] - WMn[2]*varm[2,3] -
WMn[3]*varm[1,3] + 2*varm[1,1] * (varm[1,2] + WMn[2]*WMn[3]) +
2*WMn[2]^2*varm[1,2]

##  expectation and variance assuming COVARIANCE of the numerator
##  and denominator equal to cvr, and the estimated covariance
```

```
cvr _ c((-0.1+seq(11)/10)*sqrt(nuvar*devar),covr)

ebet _ nume/deno + cvr/deno^2 - devar*nume/deno^3

vbet _ nuvar/deno^2 + devar*nume^2/deno^4 - 2*cvr*nume/deno^3

##  residual variance

##  numerator

u1 _ nuvar + (WMn[5] - WMn[2]*WMn[3] - varm[1,2])^2

##  denominator

u2 _ WMn[4] - varm[1,1] - WMn[2]^2

##  variance of numerator is   3*nuvar^2
##  variance of denominator is   devar

cvr _ c((-0.1+seq(11)/10)*sqrt(3*devar)*nuvar,covr*sqrt(3*nuvar))

rsig _ WMn[6] + varm[2,2] - WMn[3]^2 -
u1/u2*(1 + devar/u2^2) + cvr/u2^2


paste("covariance  ",covr)

paste("correlation  ",crre)
"Done"


##  sampling variance of the numerator and denominator

nuvar;devar;

##   estimates and standard errors

rout _ rbind(WMn[3]-ebet*WMn[2],ebet,vbet,rsig,covr/
```

```
sqrt(nuvar*devar))
dimnames(rout) _ list(c("Intercept","Slope","St.Err.Sl.",
"Res.var.","Covariance"),c(paste("Correlation",
(seq(11)-1),"/10"),"Est.corr."))
Mfit[[i]] _ rout
}

Mfit

## Summarize
MfitS _ matrix(0,5,12)
mns _ matrix(0,lp,12)

for (i in Ip)
{
MfitS _ MfitS + Mfit[[i]]
mns[i,] _ Mfit[[i]][2,]
}

MfitS _ MfitS/lp
MfitS[3,] _ sqrt(MfitS[3,] + apply(mns,2,var)*(1+1/lp))

MfitS
```

References

Becker, R.A., Chambers, J.M., and Wilks, A.R. (1988). *The New S Language. A Programming Environment for Data Analysis and Graphics*. Wadsworth & Brooks/Cole Computer Science Series, Pacific Grove, CA. Harville, D.A. (1974). Bayesian inference for variance components using only error contrasts. *Biometrika*, **61**, 383–385.

Johnson, E.G. and Allen, N.L. (1992). The 1990 NAEP Technical Report. Educational Testing Service, Princeton, NJ.

Johnson, E.G. and Rust, K. (1992). Population inferences and variance estimation for NAEP data. *Journal of Educational Statistics* **17**, 175–190.

Koffler, S. (1991). *The Technical Report of NAEP's 1990 Trial State Assessment Program*. Report No. 21-ST-01, NAEP. National Center for Educational Statistics, Washington, D.C., and Educational Testing Service, Princeton, NJ.

Little, R.J.A. and Rubin, D.B. (1987). *Statistical Analysis with Missing Data*. John Wiley and Sons, New York, NY.

Lord, F.M. (1980). *Applications of Item Response Theory to Practical Testing Problems*. Lawrence Erlbaum Associates, Hillsdale, NJ.

Mislevy, R.J. (1984). Estimation of latent group effects. *Journal of the American Statistical Association* **80**, 993-997.

Mislevy, R.J. (1985). Estimating latent distributions. *Psychometrika* **49**, 359–381.

Mislevy, R. and Bock, R.D. (1982). *BILOG: Item Analysis and Test Scoring with Binary Logistic Models* (computer program). Scientific Software, Mooresville, IN.

Mislevy, R.J., and Sheehan, K.M. (1989). The role of collateral information about examinees in item parameter estimation. *Psychometrika* **54**, 661–679.

Potthoff, R.F., Woodbury, M.A., and Manton, K.G. (1992). "Equivalent sample size" and "equivalent degrees of freedom" refinements for inference using survey weights under superpopulation models. *Journal of the American Statistical Association* **87**, 383–396.

Wolter, K.M. (1985). *Introduction to Variance Estimation*. John Wiley and Sons, New York, NY.

ISBN 0-16-045452-2

United States
Department of Education
Washington, DC 20208–5653

Official Business
Penalty for Private Use, $300

Postage and Fees Paid
U.S. Department of Education
Permit No. G–17

Third Class